A Computational Model of the Cognition of Tonality

Thesis

A COMPUTATIONAL MODEL OF THE COGNITION OF TONALITY

ANDREW J. MILNE, M.A., B.A. (HONS)

Submitted for the degree of Doctor of Philosophy

Faculty of Mathematics, Computing and Technology
The Open University

April 2013

Music expresses that which cannot be put into words
and that which cannot remain silent.

— Victor Hugo

Dedicated to Laurie

1959 – 1993

ABSTRACT

Tonality is the organization of pitches, both simultaneously and across time, so that certain pitches and chords are heard as attracted, in varying degrees, to other pitches and chords. Most art music from the seventeenth to the nineteenth centuries, and popular music to the present day, is heavily steeped in a musical language that makes use of tonality to define a 'central' most attractive pitch or chord called the *tonic*. It is widely thought that the feelings of expectancy and resolution induced by movements towards and away from the tonic allow composers to imbue tonal music with meaning and emotion.

In this dissertation, I identify and model some of the innate processes by which feelings of tension, resolution, stability, and so forth, are induced by successions of pitches and chords, irrespective of their harmonic consonance. By *innate*, I mean processes that do not require the learning of a musical corpus—such processes are important because they provide explanations for why tonal music, and our cognition of it, take the specific forms they do.

To do this, I introduce a novel family of mathematical methods—metrics applied to expectation tensors—for calculating the similarity of pitch collections. Importantly, such tensors can represent not just the notated pitches of tones, but also their spectral pitches (their harmonics). I then demonstrate how these techniques can be used to model participants' ratings of the fits of tones in microtonal melodies, and the fits of all twelve chromatic pitches to an established key centre (Krumhansl's probe tone data). The techniques can also be generalized to predict the tonics of any arbitrarily chosen scale—even scales with unfamiliar tunings.

In summary, I demonstrate that psychoacoustic processes, which are innate and universal, play an important role in our cognition of tonality.

PUBLICATIONS

Some ideas and figures have previously appeared in the following publications:

Milne, A. J., Sethares, W. A., and Plamondon, J. (2008). Tuning continua and keyboard layouts. *Journal of Mathematics and Music*, 2(1):1–19.

Milne, A. J. (2010). Tonal music theory: A psychoacoustic explanation? In Demorest, S. M., Morrison, S. J., and Campbell, P. S., editors, *Proceedings of the 11th International Conference on Music Perception and Cognition*, pages 597–600, University of Washington, Seattle, USA.

Milne, A. J., Prechtl, A., Laney, R., and Sharp, D. B. (2010). Spectral pitch distance & microtonal melodies. Poster presented at the International Conference for Music Perception and Cognition, University of Washington, Seattle, USA, 23–27 August 2010.

Milne, A. J., Sethares, W. A., Laney, R., and Sharp, D. B. (2010). Metrics for pitch collections. In Demorest, S. M., Morrison, S. J., and Campbell, P. S., editors, *Proceedings of the 11th International Conference on Music Perception and Cognition*, pages 77–80, University of Washington, Seattle, USA.

Milne, A. J., Sethares, W. A., Laney, R., and Sharp, D. B. (2011). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music*, 5(1):1–20.

Prechtl, A., Milne, A. J., Holland, S., Laney, R., and Sharp, D. B. (2012). A MIDI sequencer that widens access to the compositional possibilities of novel tunings. *Computer Music Journal*, 36(1):42–54.

ACKNOWLEDGMENTS

Firstly, I would like to thank my two supervisors, Dr. Robin Laney and Dr. David Sharp, for the huge amounts of time they spent reading various drafts of material as it progressed, and making numerous useful suggestions. They, correctly, insisted I presented my ideas in the simplest and most accessible form possible. Something I cannot pretend to have always achieved.

Professor Bill Sethares has been a valued colleague and friend since we wrote our first paper together in 2007. Since that time, Bill has acted as an unofficial external supervisor for both my Master's degree (completed in 2009) and over the period I have been working on this PhD dissertation. He introduced me to many key mathematical concepts, such as linear algebra and metrics, that have proved invaluable for my work. Chapter 3 draws upon a journal paper I co-authored with Bill and my two supervisors. The concepts and their development are mine—the use of Gaussian smoothing, generalizing to higher dimensions, the use of tensors, the differences between absolute and relative embeddings, and breaking the tensors into subspaces to simplify their calculation. But these all drew upon Bill's seminal input, which was the suggestion of formal metrics between vectors in the pitch domain. Bill also helped me to generalize the tensor simplifications into dimensions higher than three (above which human visualization becomes somewhat challenging). I would also like to thank Dr. Bridget Webb for helpful discussions on techniques for enumerating the subspaces of the tensors so they can be appropriately included and excluded.

Also in relation to this chapter, I would like to thank our two anonymous reviewers at the *Journal of Mathematics and Music*. Particularly the suggestions to change our matrix notation into tensor notation, and to avoid the excessive use of subscripts (something which I only had time to partially rectify in the published paper, but which I fully com-

pleted for this dissertation). I would also like to thank Professors Tuomas Eerola and Petri Toiviainen for allowing an early phase of this project to be developed as part of the Music, Mind and Technology Master's programme at the University of Jyväskylä, and Margo Schulter for setting an early challenge, the solution to which provided some important insights.

Section 2.3 and Chapter 4 draw heavily upon a paper I co-authored with my supervisors for the journal *Music Perception*. I would like to thank the three anonymous reviewers, all of whom made extremely helpful suggestions for improving the work, and Jim Plamondon—my isomorphic co-conspirator—who also gave me many useful comments on these two sections. I would like to thank Stefan Kreitmayer for assistance with the JavaScript parts of the Max patch used in the experiment, Professor Paul Garthwaite for advice on appropriate statistical tests, and my participants for taking time out of their busy schedules to listen to weird musical stimuli, and rate them.

I would like to express my profound gratitude to Dr. Simon Holland, mostly for being Simon. But, more specifically, for being an ever stimulating and fun sounding-board, inspiring novel avenues of thought, and for his enthusiasm and interest in my work. I would also like to thank all the other members of The Stern Brocot Band—Simon Holland, Vassilis Angelis, Maxime Canelli, Anthony Prechtl, and Simon Rolph—for allowing some progressive music-making to sneak its way into my allotted research time. Without creativity, it's all just numbers!

And the great and irreplaceable constants throughout this process—who deserve the most thanks—are Christine and Astra for keeping life 'normal' and creating a happy family environment in the face of my long hours of work and neurotic obsessing with finding simplifications of operations applied to multi-dimensional arrays of numbers.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Imagine listening (or actually listen) to J. S. Bach's first Prelude in C major (BWV 846). I choose this piece because it is well known, and it clearly exemplifies many of the important characteristics of tonal music that I discuss in this dissertation (including present day tonal music). The first four bars produce a gentle wave of tension and release, most noticeable with the resolution of the chord in the third bar to the chord in the fourth bar (each bar, in this piece, consists of a single arpeggiated chord). In the fifth bar, the music seems to start afresh and take a new journey for the next few bars until, in bar 11, there is a feeling of a somewhat temporary resolution—not a final destination, but a brief resting place. In bar 12, the tension increases considerably, and this is released—to some extent—in the subsequent bar. This two-bar pattern of tension (even-numbered bar) then partial release (odd-numbered bar) repeats up to bar 23 where the tension is increased further. This tension is sustained at a high overall level, occasionally peaking, until the final resolution to Cmaj in the 36th bar, which provides a strong feeling of release, resolution, and closure. This is a crude analysis and the details may differ for different listeners, but I use it to show how feelings such as tension, release, motion, rest, stability, and closure, and so forth, are typical responses to tonal music.

The primary research question addressed by this dissertation is: Can we identify, and model, the causal processes by which feelings of tension, resolution, stability, and so forth, are induced by successions of pitches and chords?

This research question is relevant to Western tonal music because, in such music, pitches are organized across time (as melodies and chord progressions), and the feelings of tension and resolution this induces are amongst tonal music's most important perceptual characteristics.

Indeed, the feelings of expectancy and fulfillment aroused by tense chords resolving to stable chords—or the lack of fulfilment when they do not—is one of the principal means by which music communicates emotion and gains meaning (Meyer, 1956; Huron, 2006).

For example, the chord progression Fmaj–Gmaj sets up a powerful expectancy that the following chord will be Cmaj: finishing on this chord produces a strong sense of resolution and closure. Moving, instead, to Amin has a somewhat surprising but pleasing effect: there is some sense of resolution but it is not as strong. Moving to F♯maj, however, gives a quite different effect: it sounds genuinely surprising, even somewhat clumsy.

Inevitably, there are many possibles causes of such responses to music. I restrict my focus in two ways. Firstly, I concentrate on those causes that are not learned; that is, 'hard-wired' or innate responses to physical stimuli. This is in contrast to Meyer and Huron and many contemporary researchers (e.g., Krumhansl (1990), Lerdahl (2001), Pearce and Wiggins (2006)) who argue that tonal expectancy is down to no more than familiarity—the learning of common patterns. Instead, I propose that innate perceptual and cognitive mechanisms play an important additional role, and one that is actually more fundamental because they may underlie music perception across cultural boundaries and are able to make predictions about which musical forms are more likely to arise. This is discussed, in depth, in Section 2.3.

Secondly, I consider only those processes that are not a function of changing levels of consonance and dissonance. For example, consider moving from the harmonically dissonant chord G7 to the consonant chord Cmaj—such a progression induces a powerful feeling of closure, and it is plausible that this is due to the transition from dissonance to consonance. However, there is clearly more to it than just this; for instance, consider the progression G7–Fmaj, which similarly transitions from dissonance to consonance, but induces little feeling of closure. Or, consider the above-mentioned progression Fmaj–Gmaj–

Cmaj where all the chords are equally consonant, but which does induce a strong feeling of closure.

By constraining my research area in these two ways, I can state my aim precisely: I am seeking to identify and model the innate processes by which feelings of tension, resolution, stability, and so forth, are induced by successions of pitches and chords, irrespective of their harmonic consonance.

In Chapter 2, I consider the overall background to this research, and I use existing literature to explain the development of my theory, and to help shape the definitions I use throughout this dissertation. Firstly, I consider various common meanings of the word *tonality*, and introduce the definition I use for this dissertation: *tonality* is the organization of pitches, both simultaneously and across time, so that certain pitches or chords are heard as attracted, in varying degrees, to other pitches and chords. I also describe how tonality induces feelings described by terms like *affinity*, *tension*, *activity*, *expectancy*, *resolution*, *stability*, and so forth. In Section 2.1, I discuss the importance of tonality—notably the manipulation of expectancy and resolution—in communicating meaning and emotion in music. In Section 2.2, I outline a broad strategy by which modelling the cognition of tonality may be approached. I hypothesize tonality is due, in part, to three mechanisms: familiarity, consonance, and affinity (*affinity* is the melodic analogue of consonance; the extent to which sequential pitches or chords are perceived to fit). Given an overall context of pitches (such as a previously established scale), we expect pitches or chords that are unfamiliar, dissonant, and have low affinity to move to pitches or chords that are familiar, consonant, and have high affinity. My focus is on modelling innate causes of affinity, and using these to model tonal cognition. I also discuss the broad methodological framework of this research: my aim is to produce parsimonious mathematical (computational) models of perception tested against experimentally obtained data. I also seek models with explanatory as well as predictive power—as explained in Section 2.3.

Section 2.3 is a more substantial section in which I categorize—in a novel way—the different processes by which physical stimuli can induce subjective mental states (sensations, feelings, and concepts), and discuss the complex interlinked causal roles played by each of these processes. It is in this section that I discriminate between the predictive and explanatory powers of models, identify the circular causal processes that operate between the perceptions of listeners and composers and the musical repertoire they create, and argue that only bottom-up models of innate processes are capable of providing effective explanations for tonal cognition. In Section 2.4, I briefly review existing theories and models of tonal cognition (I provide more focused reviews in later chapters). In Section 2.5, I give an overview of the models I have developed to meet the above research aim, and how they are empirically tested.

In Chapter 3, I introduce the general mathematical techniques that underlie all of the more specific models used in subsequent chapters. These constitute a novel family of mathematical methods for calculating the similarity of pitch collections (Secs. 3.2–3.5). The similarity values are derived by standard metrics applied to pairs of *expectation tensors* (a tensor is also known as a multi-way array), which embed either pitches or pitch classes.[1] Furthermore, these pitches (or pitch classes) can be either absolute or relative: in the latter case, embeddings of pitch collections that differ only by transposition are identical and so have zero distance; a useful feature that relates similarity to structure. Moreover, tensors of any order (e.g., order-1 vectors, order-2 matrices, order-3 three-dimensional arrays, etc.) can be formed, allowing the embeddings to reflect the (absolute or relative) monad, dyad, triad, and so forth, content of the pitch—or pitch class—collection. These embeddings also incorporate, in a principled way, the basic psychoacoustic property of pitch uncertainty by 'smearing' each pitch over a range of

1 A *pitch* is the auditory attribute associated with a specific frequency. Musical pitches can be notated with letter names and numbers like A4 (which corresponds to 440 Hz), A5 (which corresponds to 880 Hz), and so forth. A *pitch class* is an equivalence class for all pitches an octave apart. For example, the pitches C1, C2, C3, C4, C5, and so on, all belong to the pitch class C.

possible values, and the width of the smearing can be related to experimentally determined frequency difference limens, or fitted to experimental data.

The different expectation tensors provide a unified framework for representing perceptually relevant features of pitch collections, and are a generalization of a number of conventional embeddings used in music theory, including pitch vectors, interval vectors, and subset-class vectors. In Section 3.6, I demonstrate a number of applications of the expectation tensors, including some of the higher-order tensors. However, for the remainder of the dissertation, I focus my attention on the order-1 tensors—vectors—which can be used to embed spectral pitches (these are denoted *spectral pitch vectors*). *Spectral pitches* correspond to the *partials* (frequency components) of a tone—when an instrument plays a single notated pitch, it actually produces a large number of partials typically at harmonics above the notated fundamental pitch. The similarity of any two such vectors (their *spectral pitch similarity*) is used to model the affinity of the notated pitches or chords they embed.

In Chapter 4, I tackle the issue of nature versus nurture head-on. I describe a spectral pitch similarity model and an experiment designed specifically to test whether innate processes have a meaningful impact on our perception of the fit or affinity of successively played (i.e., melodic) pitches. In order to do this, I conduct a music perception experiment using unfamiliar microtonal stimuli to minimize the impact of learned responses.[2] The use of microtonal stimuli to probe into the innate processes that may underlie music perception is, as far as I know, entirely novel. The experimental data (obtained from 44 participants) indicates that innate processes, as modelled by spectral pitch similarity, do indeed play a meaningful role in our perception of melodic affinity. Furthermore, this work validates the use of microtonal stimuli as a useful and novel experimental tool for music psychologists.

Having established that spectral pitch similarity can effectively model innate perception, in Chapter 5, I introduce a related bottom-

2 *Microtonal* music contains intervals that do not occur in the familiar 12-tone equal temperament—some of its pitches 'fall in the cracks' between a piano's keys.

up model for an important and well-known set of experimentally obtained data—the probe tone data—collected by Krumhansl and Kessler (1982). These data measure the perceived level of fit of each chromatic degree with respect to a previously established tonal centre. My model provides an extremely good fit to the data and provides a plausible bottom-up mechanism for it. I then extend the same model to predict the tonic triad (the best fitting major or minor chord) of any given scale. The model's predictions accord with conventional music theory. I finish this chapter by showing how the model can be extended to predict the tonics of a variety of unfamiliar microtonal scales. I do not, at this stage, offer any empirical validation for these predictions, but suggest that doing so will provide an important test for my model.

In the final chapter—Chapter 6—I provide a critical analysis of the work, summarize the conclusions that can be made, and delineate future work that may extend what has already been achieved.

2

MODELLING THE COGNITION OF TONALITY

As outlined by Dahlhaus (1980) and Hyer (2001, 2002), the term *tonality* has a variety of related meanings, some broad and generic, others more specific. For instance, it can be used in a broad sense to describe 'the systematic organization of pitch phenomena in both Western and non-Western music', and 'a rational and self-contained arrangement of musical phenomena' (Hyer, 2002, p. 727). Or, it can take a slightly narrower definition that specifies a particular such arrangement and denotes 'a system of relationships between pitches having a "tonic" or central pitch as its most important element' (Dahlhaus, 1980, p. 52). In this latter definition, tonality refers to a specific systematic organization of pitches that was developed in the early seventeenth century, and may be contrasted with *modality*, which came before, and *atonality*, which came after.[1] It can also take an even narrower definition, where it is used as a synonym for *key*; as in, 'this piece is in the tonality of A♭ major'. As suggested by both Dahlhaus and Hyer, there is also some ambiguity about whether *tonality* refers to the organization of pitches, to the music that results from this organization, or to the feelings these organized pitches induce.

The term *tonality* was popularized by Fétis in the early nineteenth century, and he provided a useful, and suitably broad definition. He conceived of it as the sum total of the forces of attraction between successive or simultaneous notes of a scale (Hyer, 2002). For this dissertation, I use a related definition. I define *tonality* as the organization of pitches, both simultaneously and across time, so that certain pitches or chords are heard as attracted, in varying degrees, to other pitches and

1 *Modality* is the musical system existing prior to the seventeenth century whose organization is based more upon melodic principles than chordal (Dahlhaus, 1990), and which does not have a strongly defined tonic. *Atonality* refers to those systems of music developed in the twentieth century (notably serialism), which deliberately avoid structures that generate a tonic.

chords. Tonality may also induce feelings that certain chords or pitches feel tense, unstable, yearning, expectant, while others sound relaxed, stable, resolved, and fulfilled, and so forth. A simple example of this is how the seventh degree of the major scale (the so-called 'leading tone') is typically heard as being attracted to the first degree (e.g., in the C major scale, the pitch B strives towards the pitch C).

In line with common usage, I use *tonality* to refer to the organization of pitches, the resulting music, and the feelings it induces. For example, a piece of music in which the identity of the tonic is vague, may be described as having an ambiguous tonality (*tonic* is fully defined in the next paragraph). Here, the tonality that is being referred to is the effect that this music induces in the mind, not the musical structure itself.

If the sum total of attractions induced by a tonality points towards a single pitch or chord, this is called the *tonic*. From the start of the seventeenth to the end of the nineteenth centuries, Western music has favoured structures that induce a single tonic over long periods of time (and such music is still commonplace today). This type of music is called *major-minor tonality*, *harmonic tonality*, *common-practice tonality*, or *tonal-harmonic music*;[2] it makes use of major and minor scales and triads, and cadences such as V^7–I, IV–V–I, and II–V–I, to induce a strong and unambiguous tonic pitch and triad on scale degree $\widehat{1}$.[3] In major-minor tonality, harmonic cadences—such as the above—are a vital structural component because they strongly define the tonic. Indeed, Lowinsky (1961) argued the use of such cadences was what transmuted medieval modal music into fully tonal music with a consolidated tonic, and coined the memorable phrase: 'the cadence is the cradle of tonality' (p. 4).

However, other types of tonality are possible. For example, Indian classical music establishes a strong tonic not with cadential chord pro-

2 These terms are broadly synonymous, but have slightly different connotations. For example, *major-minor tonality* emphasizes the use of the major and minor scales rather than the medieval modes to induce a tonic, *harmonic tonality* and *tonal-harmonic* emphasize the use of chords, while *common-practice tonality* emphasizes a specific historical period.

3 Ordinal scale degrees are denoted with Arabic numerals with a caret above—the first scale degree usually corresponds to the tonic of the scale. Roman numerals indicate triads, the value referring to the scale degree of the chord's root.

gressions but by using a constant drone above which the melody plays (Jairazbhoy, 1995; Widdess, 1981); and atonal music can hint at rapid successions of different and antagonistic tonics, none of which it resolves to. Indeed, Schoenberg (1978)—the originator of the serial technique of atonal composition—objected to the term *atonal*:

> Everything implied by a series of tones constitutes tonality, whether it be brought together by means of reference to a single fundamental or by more complicated connections. That from this single correct definition no reasonable opposite corresponding to the word 'atonality' can be formed, must be evident. (p. 432)

Furthermore, it is possible that microtonal scales—those built from intervals not found in twelve-tone equal temperament—may produce musical structures quite different to those found in common-practice tonality, but which still support a tonic; I explore this theoretical possibility in Section 5.

My broad definition of *tonality* allows the term to encompass all these different types of music, while I use *major–minor tonality*, and so forth, to refer to the more specific form of Western music that originated around the start of the seventeenth century.

2.1 WHY IS TONALITY IMPORTANT?

Tonality matters because it is one of the principal means by which music can communicate emotions. There are many emotional-sounding adjectives that are used to describe the effects of tonality. For example, to describe the feelings induced by tonality, Fétis (1844) utilizes the words *tendance* ('tendency'), *appeler* ('summon'), *résolution* ('resolution'), *anxiété* ('anxiety'), *repos* ('repose'), *agitations*, *crises nerveuses* ('nervous crises'), *plaisir* ('pleasure'), and so forth. And words like *yearning*, *desire*, *seeking*, *expectant*, *fulfilled*, *away*, *home*, *foreign*, *domestic*, *tense*, *relaxed*, *sour*, *sweet*, *stability*, *instability*, *activity*, *motion*, *rest*, *suspense*, *anticipation*, *resolution*, *closure*, *completion*, *surprise*, *deceptive*, *incomplete*, *inter-*

rupted, and so on, are commonly used descriptors (e.g., Huron (2006, Tables 9.1, 14.1, & 14.2) compiles an extensive list of adjectives, given by listeners, to describe the feelings induced by different scale degrees and chromatic chords).

It seems the effects produced by tonality are redolent with broad emotional and conceptual associations. Considered in this way, tonality comprises an evocative set of signs that gives music considerable communicative potential.[4] This tallies with the way music seems to function as a language that is able to communicate broad emotional states, with great depth, but with limited precision. Citing a number of studies, Juslin and Laukka (2004, p. 219) write: 'there is usually high agreement among listeners about the broad emotional category expressed by the music, but less agreement concerning the nuances within this category'.

Furthermore, as suggested by Hyer (2002), the hierarchical structure of major-minor tonality—in which there is an unambiguous central tonic, which 'controls' all other pitches and chords—mirrors the hierarchical societies within which this musical system gained ascendancy. Indeed, it could be argued that the rejection of major-minor tonality in Western art music of the twentieth century betokens the revolutionary impulses of that period.

At the very heart of tonality is attraction or, put differently, *expectancy*—the perceived expectation that a pitch or chord will resolve to another pitch or chord. Meyer (1956) famously argued that it is the manipulation of expectancy that gives music meaning and emotion—a composer might fulfill our expectations, or surprise us, or introduce a tension-inducing delay before fulfilling them. By 'orchestrating' these musical devices, a composer can evoke a broad range of emotions.

Recently, Huron (2006) has provided a comprehensive theory to explain precisely how expectancies—and their violations and fulfillments—can account for the emotions they induce, and why such

4 I use *sign*, in the semiotic sense, to mean something that refers to something other than itself (Chandler, 2002).

emotions might serve an evolutionary purpose.[5] The theory is summarized by the acronym ITPRA, which stands for five 'functionally distinct physiological systems' (p. 7): *imagination*, *tension*, *prediction*, *reaction*, and *appraisal*.

Huron claims that positively and negatively valenced emotions, respectively, encourage adaptive behaviours and discourage maladaptive behaviours (hence they serve an evolutionary function). And that emotions related to expectancy function similarly to encourage accurate predictions and appropriate preparation. In this way, the manipulation of musical expectancy can tap into primordial emotions 'including surprise, awe, "chills," comfort, and even laughter' (Huron, 2006, p. 4).

The *imagination response* relates to the emotional content of imagined situations. For example, in imagining a specific future scenario we not only think through its consequences, we also—to some extent—feel them. It is through this mechanism that we are able to anticipate the emotional implications of future states.

The *tension response* is the increase in arousal and attention produced by the anticipation of an outcome. Its associated emotions are somewhat negatively valenced because increased arousal and attention require more energy.

The *prediction response* occurs directly after an event. Given an outcome, it rewards or punishes according to the accuracy of the previously made prediction of the outcome. So long as an outcome is accurately predicted, this can result in positively valenced feelings even when that outcome is inherently bad.

The *reaction response* is the immediate, and unconscious, assessment of the event itself. The resulting emotions are typically produced by a violation of a well-learned schema; that is, surprise. The reaction response is negative in valence and, in music, serves an important role by creating a *contrastive valence* (discussed below) that exaggerates the

5 Both Meyer and Huron use the term *expectation* more often than *expectancy*. In this dissertation, I use *expectation* for the expected value of a random variable–or some function, thereof–and where possible I use *expectancy* for the subjective feeling of expecting something.

effects of the more neutral or positive conscious appraisal that occurs later.

The *appraisal response* is the more slowly developing conscious assessment of the outcome. It may draw upon 'complex social and contextual factors' (Huron, 2006, p. 15)

Because the five systems are functionally distinct they may produce consistent or inconsistent responses. This allows for an important type of interaction between the different response systems—*contrastive valence*—which is that the valence of an outcome, as determined by appraisal, is amplified by the degree to which it was previously oppositely valenced. Typically, the appraisal of a musical outcome is neutral or positive (because it produces little real threat), so any positivity is enhanced when this outcome is preceded by events that increase tension, incorrect identification, and surprise. In other words, we are more pleased by an unexpected positive outcome than an expected positive outcome. For example, the pleasure induced by a consonant chord is enhanced when it follows a long sequence of dissonant chords.

With respect to explaining how tonality can evoke emotions, the most relevant responses are those of tension, prediction, reaction, and appraisal. The tension response, and its associated emotions, are tonally triggered by creating a strong expectancy—for example, a V^7 chord creates a strong expectancy of a proceeding I (or i, depending on the context). And tension can be increased when the expected resolution is delayed, for example by staying on the V^7 for longer than expected.

The prediction response, which produces positively valenced feelings, such as pleasure, and negatively valenced feelings, such as weirdness, is triggered by organizing pitches—vertically and horizontally—in a way that, respectively, conforms with expectation, or does not.

Huron argues that musical surprise (which represents a failure of prediction), produces negatively valenced prediction and reaction responses, but a positive (or neutral) appraisal response, and that this can induce three emotions: frisson, laughter, and awe. He considers these to be related, in respective order, to the three common responses to sur-

prise: fight, flight, and freeze. The fight response is characterized by an aggressive posture and this, in turn, is related to the piloerection (hairs standing up) that is a characteristic of frisson ('the hairs on the back of my neck are standing up'). The flight response is characterized by rapid respiration, which is related to the modified panting that characterizes laughter. The freeze response is characterized by the holding of breath, which is related to the gasp associated with awe. These feelings are amplified by the contrastive valence between the negative feelings associated with surprise (due to the prediction and reaction responses) and the appraisal response that 'realizes' there is no real danger. The initial fear produced by the reaction response is short-lived and is not conscious, but the resulting emotion—frisson, laughter, or awe—is.

In this way, Huron argues that there are deeply embedded psychological processes that can account for how the manipulation of pitch-based expectation or attraction—that is, tonality—can induce a variety of emotional states. For Huron, expectancy is solely down to learning: We expect pitch x to sound at the same time as pitches y and z because we have heard that chord many times before; we expect chord x to be followed by chord y then chord z, because we have heard that progression many times (both in general music and in the piece itself).[6]

However, I argue in the remaining sections of this chapter that there are plausible, and experimentally demonstrable, innate processes that can account for certain important aspects of tonal expectancy (attraction).

2.2 A STRATEGY FOR MODELLING TONALITY

It is a plausible hypothesis that tonal attraction is due, in part, to a listener wanting a transition from an 'unpleasurable' state to a 'pleasurable' state. For instance, we might class the dominant seventh chord V^7

6 Huron's disavowal of psychoacoustic explanations in Huron (2006) is somewhat surprising considering much of his seminal earlier work where he provides psychoacoustic and other bottom-up explanations for a number of related phenomena such as the choice of intervals favoured in two-part polyphony, common scales in world music, and the rules of Western voice-leading (Huron, 1991, 1994, 2001).

as unpleasurable, and the tonic chord I as pleasurable; hence we hear the former as attracted to the latter. As mentioned above, it is also plausible that attraction is due, in part, to a listener becoming accustomed to certain melodic or harmonic moves. For example, we may have become familiar with music that uses the progression V^7–I, so we expect to hear the former chord move to the latter.

This gives clues as to how to model tonal attraction. What possible mental processes might we use that could be associated with pleasure/displeasure, and so forth? Perhaps the most obvious candidates are familiarity/unfamiliarity, sensory consonance/dissonance, and sensory affinity/disaffinity.

Familiarity/unfamiliarity is here defined as the extent to which a listener is accustomed to a given musical event in a given context. It is a function of that event and all previous musical events (and their contexts) heard by the listener. The familiarity of an event may or may not be directly perceived. For example, major triads are commonly used in Western music so, to a Westerner, they will be familiar. Similarly, the chord progression iii–I is rare in Western music so it will be unfamiliar to a Western listener. Familiar musical stimuli are typically regarded as more pleasurable (e.g., North and Hargreaves (1995) found a correlation of $r(58) = .91$ between 'liking' and 'familiarity' ratings of sixty musical excerpts averaged over 25 participants).[7] Under Huron's ITPRA theory, this is because the perception of familiar events is rewarded by the prediction response.

Consonance/dissonance is here defined as a perceptual/cognitive attribute that quantifies the degree of pleasure, fit, or compatibility of simultaneous tones as a function of their pitch (fundamental frequency) and timbre (spectral content), but irrespective of their context within a progression of chords or pitches. In this sense, consonance/dissonance is defined as a purely vertical phenomenon (*vertical* in the sense of musical notation) and is the same as Tenney's second consonance-dissonance concept (CDC-2) (Tenney, 1988). *Sensory consonance* is that part of con-

7 In this dissertation, I follow APA (American Psychological Association) guidelines for statistical reporting and enclose the degrees of freedom inside parentheses.

sonance/dissonance that is down to psychoacoustic phenomena, such as *partials* that are close enough in frequency to create *roughness* (unpleasant beating), or partials that fall into a harmonic series creating a pleasant sensation of *toneness* or *fusion*.[8] The non-sensory part of consonance/dissonance is down to learning—that is, as we become familiar with any given chord it becomes more consonant (pleasurable)—and is, therefore, included in the above familiarity/unfamiliarity category. All these aspects of consonance/dissonance are examined in greater depth in the introduction of Chapter 4, and suitable models are discussed in Section 4.1.2.

Affinity/disaffinity is here defined as a perceptual/cognitive attribute that quantifies the degree of pleasure, fit, or compatibility of non-simultaneous tones as a function of their pitch (fundamental frequency) and timbre (spectral content). It can also be thought of as the melodic analogue of consonance/dissonance, and is equivalent to Tenney's first consonance-dissonance concept (CDC-1) (Tenney, 1988). In this sense, affinity/disaffinity is defined as a purely *horizontal* phenomenon (horizontal in the sense of musical notation). *Sensory affinity* is that part of affinity/disaffinity that is down to psychoacoustic phenomena such as *spectral pitch similarity*, which is a measure of the similarity, in pitch, of the partials of two tones or chords (I give a fuller definition in Sec. 2.5). The non-sensory part of affinity/disaffinity is down to learning—that is, as we become familiar with any succession of pitches or chords we perceive them to have greater affinity (the succession is more pleasurable)—and is, therefore, included in the above familiarity/unfamiliarity category. Models of affinity are examined in more depth in Section 4.1.1.

In this way, we can argue that a chord or pitch that is unfamiliar, dissonant and poor-fitting, will be attracted to resolve to a chord or pitch that is familiar, consonant, and good-fitting. Familiarity is, by

8 *Partials*—also known as *overtones*—are the frequency components (sine waves) that make up a complex tone. *Complex tones* comprise a number of partials at a variety of frequencies and phases. *Harmonic complex tones* contain partials whose frequencies are all at (or close to) integer multiples of a single *fundamental* frequency. Most pitched Western musical instruments produce harmonic complex tones containing numerous partials.

definition, a top-down or learned process. Consonance and affinity are, in part, bottom-up (because the consonance and affinity of different intervals, chords, and chord progressions, can be plausibly modelled without any statistical analysis of their prevalence in a relevant musical corpus).

In this dissertation, I focus on sensory affinity rather than sensory consonance/dissonance—that is, on the horizontal rather than vertical aspects of tonality. The reason for this is because the latter has already been widely studied, but also because it is not able to provide an explanation for basic tonal effects. For instance, the chord progression Fmaj–Gmaj–Cmaj induces a powerful feeling of attraction from the second chord to the third chord. Yet, in this example, the three chords are equally consonant (they are all major triads). Clearly, it is the overall context, which is established across time by the three chords, that determines each chord's tonal effect; it is not their individual levels of consonance/dissonance.

I also focus on *bottom-up models*—models of innate perceptual/cognitive processes that do not rely on statistical analyses of a musical corpus. The reason for this—as I argue in Section 2.3—is because only bottom-up models can provide explanations for why music takes on certain specific forms rather than others.

2.2.1 *Methodology*

Before proceeding to this, however, I feel it will be helpful briefly to outline the methodological framework of this research. I take an empirical approach in which computational models of perception and cognition are tested against experimentally obtained data. In an experiment, we know the values of a set of *independent variables* (or just *variables*) and we measure the values of a set of *dependent variables* that occur over different values of the independent variables. The resulting values constitute the experimental *data*. For instance, we may ask participants to rate how well pitches at different intervals 'fit' together—here, the

independent variable is the interval used, and the dependent variable is rating of 'fit'.

A *numerical*, *mathematical*, or *computational model* is a mathematical function whose inputs are the values of the independent variables used in the experiment and a set of *parameters*. The parameters' values are held constant over all values of the independent variables, and the model's output values are compared with the experimental data to obtain a *discrepancy* value, which measures how well the model fits the data. The discrepancy is minimized by optimizing the values of the parameters. In this way, different parameter values can be thought of as indexing through different models until the best-fitting model is found.

I seek to obtain a good fit with the experimental data using the most parsimonious model possible. A *parsimonious* model is one that is inflexible: across the whole space of parameter values, the resulting model produces values that are close to only a relatively small range of possible data (Pitt et al., 2002; Honing, 2006).

An advantage of a good-fitting parsimonious model over a good-fitting non-parsimonious model is that only the former generalizes beyond the specific sample of data to which they are fitted. This is because it models the underlying deterministic process (whose effect is constant across different data samples) rather than the stochastic noise in the data (whose effect is not constant across different data samples). For instance, when a non-parsimonious model is optimized to a noisy data set, it may fit better than a parsimonious model; but, when compared with different data samples, it will fit worse than the parsimonious model. Over a number of data samples, the parsimonious model will, on average, fit better (Zucchini, 2000).

In Chapters 4 and 5, I use k-fold cross-validation to ensure the models I test are able to generalize sufficiently (i.e., they are parsimonious) and are fairly compared.[9] And, as mentioned above—and fully explored in

9 I use cross-validation rather than methods such as BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) because these are sensitive to only one aspect of model flexibility—their number of parameters—and insensitive to their functional form (Pitt et al., 2002). This can be an important factor when comparing

the next section—I also favour models that are able to provide effective explanations as well as effective predictions.

With respect to input variables, my models use symbolic representations of music rather than live audio input. For example, when modelling a perceptual or cognitive response to a chord, I may represent it with the frequencies and amplitudes of the partials that one would typically find when that chord is played by a familiar Western instrument (e.g., wind, string, or voice) rather than precisely measuring the spectral content of the stimulus. The principal reason for this is because my interest is not in precisely modelling responses to specific musical events; rather, it is in modelling responses to general musical events that are applicable across a wide range of musical instruments, environments, and musical contexts.[10] For example, I would not model the precise differences between the chord progression Fmaj–Emaj–Dmaj played by a string quartet in a recital hall and the same progression played by an acoustic guitar in someone's front room. Instead, I would model the differences between this chord progression and a different chord progression such as Fmaj–Gmaj–Emaj (regardless of the precise instrument used, and its location).

2.3 MODELS OF MENTAL STATES: NATURE AND NURTURE

In this section, I argue that bottom-up models of innate processes (nature) have greater explanatory power than top-down models of learning processes (nurture). I also discuss the difficulty faced by researchers who wish to establish the extents to which an experimentally observed

nonlinear models, such as my own, with linear models. Also, for my models, MDL (Minimum Description Length) cannot be derived analytically, and numerical calculations would be prohibitively time-consuming. Because k-fold cross-validation can be calculated within a feasible time-period, it is the most useful method. For linear models, k-fold cross-validations with appropriate values of k are asymptotically equivalent to AIC or BIC (Stone, 1977; Shao, 1997), so these are related techniques.

10 Other reasons are that the resulting models are simpler to build; the human auditory system is able to perform important tasks—such as polyphonic pitch detection—that we cannot yet match with computational models, but which are trivial with symbolic data; spectral representations of the stimuli can be easily transformed by psychoacoustic parameters whose values are unknown a priori.

effect is due to top-down and/or bottom-up processes, and propose a solution: one that I utilize in Chapter 4.

In order to do this, in Section 2.3.1, I discuss the purpose of modelling and define *explanation* and *prediction*. In Section 2.3.2, I provide a novel categorization of the processes that can convert a physical stimulus into a mental response, thereby providing a precise definition of nature and nurture processes, and bottom-up and top-down models thereof. In Section 2.3.3, I highlight how composers create a causal loop between processes of music perception and the musical repertoire, and the implications of this loop for the explanatory power of top-down and bottom-up models. In Section 2.3.4, I demonstrate the importance of exploring causal assumptions before selecting between models—I highlight two examples where ignorance of the above-described causal loop has resulted in bottom-up models being unjustifiably rejected solely on the grounds of their lower predictive power—and indicate some of the types of experimental intervention that can test specific casual claims. The experimental procedure utilized in Chapter 4 is an example of such an intervention: one that is designed to test the strength of a hypothesized bottom-up casual link from spectral pitch similarity to perceived affinity.

2.3.1 *Explanation and Prediction*

In the science of music perception and cognition, we aim to model subjective mental responses (sensations, feelings, and thoughts) with variables derived from musical events or associated non-musical events. That is, given the cultural and environmental experiences of a listener, we seek to *predict* their likely mental responses to a tone, or pair of tones, or chord progression, or rhythm, or spectrum, and so forth, or some combination thereof. Furthermore, we also seek to *explain* why this stimulus has this effect.

For scientific models, this latter explanatory role is arguably of greater import than the predictive role (as argued by, e.g., Deutsch

(1997), and Lewandowski and Farrell (2011)). This is because a predictive-only model (a 'black-box' model) may provide accurate predictions of the results of an experiment, but it gives little insight into the processes that lead to those results. An explanatory model (a 'white-box', or 'clear-box', model), on the other hand, provides a series of tightly defined and interrelated causal effects, each of which is plausible and somewhat inflexible.[11] The advantage of an explanatory model is that it gives us greater insight into how reality works: this enables us to use its underlying principles (laws) in areas that were not part of the original experiment or observation. In other words, the identified principles are generalizable, and hence serve as powerful tools with which to manipulate and understand nature. A predictive-only model is unlikely to provide such generality and utility.

To illustrate, Deutsch (1997, p. 4–5) asks us to

> Imagine that an extraterrestrial scientist has visited the Earth and given us an ultra-high-technology 'oracle' which can predict the outcome of any possible experiment, but provides no explanations .... How would the oracle be used in practice? In some sense it would contain the knowledge necessary to build, say, an interstellar spaceship. But how exactly would that help us to build one, or to build another oracle of the same kind—or even a better mousetrap? The oracle only predicts the outcomes of experiments. Therefore, in order to use it at all we must first know what experiments to ask it about. If we gave it the design of a spaceship, and the details of a proposed test flight, it could tell us how the spaceship would perform on such a flight. But it could not design the spaceship for us in the first place. And even if it predicted that the spaceship we had designed would explode on take-off, it could not tell us how to prevent such an explosion. That would still be for us to work out. And before we could work it out, before we could even

11 Another feature, sometimes claimed, of explanatory models is that they are understandable to humans (e.g., Trout (2007)); this is not a feature I focus on here.

> begin to improve the design in any way, we should have to understand, among other things, how the spaceship was supposed to work. Only then would we have any chance of discovering what might cause an explosion on take-off. Prediction—even perfect, universal prediction—is simply no substitute for explanation.

Busemeyer and Diederich (2010, p. 4) provide a similarly instructive, but more down-to-earth, comparison between two models of the mean response time taken to perform a complex task as a function of the number of training trials. One model is a predictive-only power-law model, the second is an explanatory memory-trace model. Only the latter model can make predictions outside the narrow scope of the experiment—for example, predicting the accuracy of responses as a function of the number of training trials (which does not follow a power law, and so cannot be predicted by the former model).

In summary, I suggest a straightforward definition: an *explanation* identifies the earliest relevant causes of the effect under consideration. The further back in the causal chain the model goes, the greater its explanatory potential; having said that, it is obviously convenient to draw a reasonable boundary beyond which stepping further back in the causal chain becomes less relevant. For example, music perception clearly relies upon acoustical events, but the precise physical mechanisms by which sound is carried through the medium of air can be glossed over and treated as an established fact.

2.3.2 *Mental Processes: Nature and Nurture*

In this section, I propose a novel categorization of mental processes that convert a physical stimulus into a mental (private, subjective) response. The processes are categorized according to the type of input variable they require. I subsequently show how this enables us to tightly de-

fine *nature* (*bottom-up*) processes and *nurture* (*top-down*) processes, and to explore the causal interactions between them.[12]

I make a fundamental distinction between *nature* processes (which do not require learning) and *nurture* processes (which do require learning), and a subsidiary distinction between *intrinsic* processes (which do not require a non-domain stimulus) and *extrinsic* processes (which do require a non-domain stimulus).[13] The terms nature, nurture, intrinsic, and extrinsic all refer to a specific domain which, in this discussion, is music (but any alternative domain could be substituted). These processes likely operate simultaneously and are causally interlinked. Together, they produce a set of sensations, feelings and concepts that can be 'observed' by introspection, and then reported as empirical data. However, I assume that, for the most part, the observer has limited insight into which process caused which feeling, or the extent to which it contributed to any specific feeling. The four types of process are illustrated as causal diagrams in Figure 2.1.

Prior to detailing each of these four processes, let me first introduce some variable names that will be used in Sections 2.3.2.1–2.3.2.4.

Let a musical event m at time t be denoted m_t, and let a non-musical event n at time t be denoted n_t. The time $t = 0$ represents the present moment so the current musical event is denoted m_0. Examples of m might be a categorical variable with values like 'major chord' and 'minor chord'; a scalar representing the pitch of a tone; a probability dis-

12 Terms such as 'nature' and 'nurture', 'innate' and 'learned', 'bottom-up' and 'top-down' are typically vaguely defined in cognitive literature, and often take on contradictory meanings. For example, for Sun (2001) 'bottom-up' refers to a process that is driven by implicit rather than explicit long-term memory, whereas in much music perception research it typically refers to sensory processes driven by psychoacoustical mechanisms (e.g., Ferguson et al. (2011)). There are a variety of other terms with meanings similar to bottom-up such as 'context-independent' (Woolhouse, 2009), 'data-driven' (Dennett, 1994), 'stimulus-driven', and 'sensory' (Bigand et al., 2003); and terms with meanings similar to 'top-down' such as 'statistical' (Pearce and Wiggins, 2006) and 'expectation-driven' (Dennett, 1994), and 'cognitive' (Bigand et al., 2003).

13 My categorization of processes is similar to Sloboda and Juslin's 'sources of emotion', but there is an important difference: their principal categorization is determined by the absence, or presence, of non-musical events (*intrinsic* versus *extrinsic*); ours by the absence, or presence, of previous (learned) musical events (*nature* versus *nurture*) (Sloboda and Juslin, 2001).

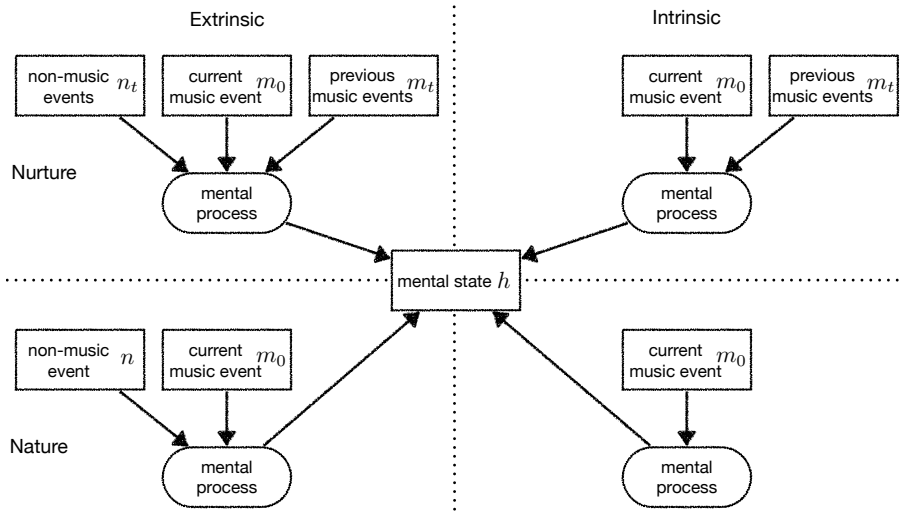


Figure 2.1: Four categories of mental processes by which physical stimuli are transduced into subjective mental states.

tribution of a feature like pitch, timbre, or loudness; a vector or array representing a set of features.

Let the mental state induced by a musical event at time t be denoted $h(m_t)$, and the mental state induced by a non-musical event at time t be denoted $h_{\mathrm{n}}(n_t)$. These might be a scalar or vector of values (e.g., a scalar representing the degree of tension felt, or a vector whose elements are the co-ordinates of an emotion space, or a categorical probability mass function over basic emotions). The function h_{n} may simply be empirical observations of the expected mental states induced by different non-musical events, rather than an explanatory model.

Let $s(x, y)$ be a measure of the similarity of x and y, which returns a value of zero when they are maximally dissimilar, and which typically takes a value of unity when they are maximally similar. For example, a very simple such measure might assign a 1 when $x = y$, and 0 when $x \neq y$.

2.3.2.1 *Extrinsic nurture processes*

Extrinsic nurture processes, in the top left corner, have three types of input variables: the *current musical event* (that which is held in short-term

memory); *previous musical events* (as stored in implicit or explicit long-term memory); *previous non-musical events*.[14]

In extrinsic nurture processes, a musical event becomes associated with a non-musical event by familiarity: listening to a piece of music, we may find that its chord progression, melodic shape, rhythmic pattern, or some other characteristic, is one that we have learned is often used in the 'sad' part of a film. In this way, we learn an association between that musical characteristic and feeling sad. It is plausible that, within a culture, some such associations may be almost universal (minor chords are sad) while others may be quite specific ('honey, they're playing our song'). This is also the principal process by which we learn the meanings of words: by repeated association with an object or concept.[15] In the case of language, the learned meanings of words can be quite precise: in music, such meanings are typically of a more vague or emotional character. In semiotics this process is termed *symbolic signification*, Sloboda and Juslin (2001) denote it *associative sources of emotion*, while this category includes Juslin and Västfjäll's (2008) mechanisms of *evaluative conditioning* (which applies to implicit long-term memory) and *episodic memory* (which applies to explicit long-term memory).

Extrinsic nurture processes are, by definition, due to learned associations and can, therefore, be mathematically modelled by a statistical analysis of associations between (co-occurrences of) specific musical events and non-musical events as experienced by an individual. Clearly, we cannot have access to an individual's experiences at every moment of his or her life, so a sample is required—a *corpus* or *training set*—of musical and non-musical variables that are likely to co-occur in his or her culture.

14 *Short-term memory* has limited capacity but can be accessed quickly. It is typically thought to decay within twenty seconds (Peterson and Peterson, 1959). Long-term memory has unlimited capacity, and lifetime duration. It can be *explicit* (episodic) allowing specific events to be consciously remembered, or it can be *implicit* (schematic) allowing general rules and patterns to be unconsciously assimilated.

15 Elements of language, particularly phonemes and graphemes, may also carry broad iconic significations (Wescott, 1971; Lansberg, 1980; Gell, 1995; Ramachandran and Hubbard, 2001), but knowing the precise meaning of most words still requires learning.

One simple way of mathematically formalizing this is the following model:

$$h(m_0) = f\left(\sum_{t \in \mathcal{T}_\mathrm{m} \cap \mathcal{T}_\mathrm{n}} h_\mathrm{n}(n_t)\, s(m_0, m_t)\right), \tag{2.1}$$

where the musical corpus contains values of m_t over $t \in \mathcal{T}_\mathrm{m}$, and the non-musical corpus contains values of n_t (or $h_\mathrm{n}(n_t)$) over $t \in \mathcal{T}_\mathrm{n}$.

The variables entering this equation are m_0, m_t, and n_t (as illustrated by the causal paths in Fig. 2.1). If we are modelling the mental state of a single participant or a group of participants with a similar cultural background, m_t and n_t can reasonably enter as constants (or parameters to be optimized). If we are modelling mental states as a function of the current musical event and the cultural background of each participant, both m_t and n_t can enter as variables.[16] The function h_n may also be modelled as being culturally determined; that is, we might expect that members of different cultures will, on average, respond to the same non-musical events in different ways.

The summation in (2.1) is a weighted sum of the mental states induced by non-musical events. Each weight is the similarity of the current musical event and the musical event that occurred simultaneously with the non-musical event. The resulting sum can be transformed into a weighted mean by dividing by $|\mathcal{T}_\mathrm{m} \cap \mathcal{T}_\mathrm{n}|$. For example, we may wish to model the happiness or sadness of a chord by how relatively often similar chords have been played simultaneously with happy and sad non-musical events.

2.3.2.2 *Intrinsic nurture processes*

Intrinsic nurture processes, in the top-right corner of Figure 2.1, have two types of input variable: the current musical event, and previous musical events. Here, we are concerned with learned musical patterns that extend in time or frequency or both. For example, due to its prevalence

16 In modelling terms, all values of m_t and n_t over t, will be adjusted *en masse* by switching between corpuses. Clearly, the corpuses should appropriately reflect the participant's cultural background.

in music, we may be familiar with the short melodic phrase (sequence of scale degrees) $\widehat{3}$–$\widehat{2}$–$\widehat{1}$ (Schenker's *Urlinie*). Upon hearing the scale degrees $\widehat{3}$–$\widehat{2}$ we may, therefore, expect to hear $\widehat{1}$ and be surprised if the following note is the comparatively rare $\sharp\widehat{4}$. Intrinsic nurture processes, therefore, provide ways in which feelings of surprise, expectancy, resolution, familiarity, alienness, and so forth can be expressed (however, I do not intend to suggest that learning is the only source of such feelings, far from it). This mechanism is termed *musical expectancy* by Juslin and Västfjäll (2008).

Intrinsic nurture processes are, by definition, due to learned patterns of musical features (across pitch, time, and timbre) and can, therefore, be mathematically modelled by a statistical analysis of the musical events experienced by an individual. As before, we cannot access an individual's experiences at every moment of his or her life, so we use a corpus of musical events that are likely to occur in his or her culture.

One simple way to mathematically formalize this is the following model:

$$h(m_0) = f\left(\sum_{t\in\mathcal{T}_\mathrm{m}} s(m_0, m_t)\right), \tag{2.2}$$

where the musical corpus contains values of m_t over $t \in \mathcal{T}_\mathrm{m}$.

The variables entering this equation are m_0 and m_t (as illustrated by the causal paths in Fig. 2.1). If we are modelling the mental state of a single participant or a group of participants with a similar cultural background, m_t can reasonably enter as constants (or parameters to be optimized). If we are modelling mental states as a function of the current musical event and the cultural background of each participant, m_t can enter as a variable.[17]

The summation in (2.2) is a weighted count of the previous musical events in a corpus. They are weighted by their similarity to the current musical event. The resulting sum can be normalized into a weighted

17 In modelling terms, all values of m_t over t, will be adjusted *en masse* by switching between corpuses. Clearly, the corpuses should appropriately reflect the participant's cultural background.

prevalence by dividing by $|\mathcal{T}_{\mathrm{m}}|$. So, if there are lots of events in the corpus similar to the current event, the resulting value will be high. The function f converts this weighted prevalence into the resulting mental state. This function may be very straightforward—we might use a linear function of the prevalence to model the perceived consonance of chords so that common chords are modelled as consonant and uncommon chords as dissonant (using a corpus appropriate to the participant). Or we may wish to model the expectancy of a chord, given the two previous chords, by the prevalence of three-chord progressions in an appropriate corpus.

2.3.2.3 *Extrinsic nature processes*

Extrinsic nature processes, in the bottom-left corner of Figure 2.1, are a function of non-musical events (past and present) and the current musical event. In these processes, a response is associated with a stimulus by similarity (not familiarity): a piece of music with a fast tempo and 'bouncy' melody can communicate excitement or arousal by analogy with bodily movement or non-verbal speech patterns, or a crack in a vocal performance may suggest the emotional fragility of the performer. It would seem that such processes can communicate a wide range of broad feelings—Juslin and Laukka (2004) provide numerous examples of musical features that carry meanings that can be associated, by similarity, with speech prosody and body movement. In semiotics these processes are termed *iconic* or *indexical signification* (Sloboda and Juslin use the related term *iconic sources of emotion*), and it includes Juslin and Västfjäll's (2008) mechanisms of *emotional contagion* and *visual imagery*.

Extrinsic nature processes are, by definition, due to perceived similarities between musical and non-musical events. This can be mathematically modelled by the mental states induced by non-musical events that are, in some sense, similar to the current musical event.

One simple mathematical formalization of this process is the following model:

$$h(m_0) = f\left(\sum_{n\in\mathcal{N}} h_{\mathrm{n}}(n)\, s(m_0, n)\right). \tag{2.3}$$

This model has a single variable m_0, with the differing values of n entering the model as constants parameterized by the range of values considered, which is denoted by the set $\mathcal{N}$. The function h_{n} may be modelled as being culturally determined; that is, members of different cultures can be modelled as, on average, responding to the same non-musical events in different ways.

The summation is a weighted sum of the mental states induced by different non-musical events. They are weighted by the similarity of each non-musical event and the current musical event. This can be transformed into a weighted mean by dividing by $|\mathcal{N}|$. So, if there are some non-musical events similar to the current musical event, and others that are dissimilar, the resulting mental state is modelled as the weighted mean of the mental states induced by the former non-musical events. For example, we may wish to model the arousal produced by a piece of music, and model certain characteristics of music (e.g., high loudness, high pitch, bright timbre, fast tempo) as being similar to the physical manifestations of arousal (e.g., fast and energetic body movements and vocalizations). If the current musical event has these properties, it will have a high similarity with these non-musical events, hence the resulting mental state is modelled as high in arousal.

2.3.2.4 *Intrinsic nature processes*

Intrinsic nature processes, in the bottom-right corner of Figure 2.1, have just one type of input variable: the current musical event. In such processes, a specific type of musical event is effectively 'hard-wired' to a mental response. At a basic level, any given stimulus is associated with a 'raw feel' or sensation—for example, the private subjective experience of pitch, timbre, or loudness. Also at a basic level, a stimulus may

induce a basic emotional response—analogously to how the smell of bread baking in an oven may induce a feeling of hunger, or the flash of a big cat's teeth may induce fear, a sudden loud sound may stimulate shock, surprise, or even pain. However, the feelings involved may not always be quite so basic. For instance, there is a well-established link between stimulus complexity and pleasure (Berlyne, 1970), and it has been demonstrated how music that is too simple for a participant is found tedious, and music that is too complex is found unpleasant, with a pleasure-maximizing 'Goldilocks' complexity somewhere in-between (Vitz, 1966; Heyduk, 1975; North and Hargreaves, 1995). It is plausible that there is an evolutionary advantage to such behaviour: it is useful for us to invest time solving (solvable) problems, but also useful that we should avoid situations where we cannot properly resolve, or make out, our percepts (such as venturing into the dark woods at night). This process includes Juslin and Västfjäll's (2008) mechanism of *brain stem reflex*.

A possible concrete example of this is the displeasure associated with 'dissonant' chords that are *rough* (due to the rapid beating of partials that are close in frequency) and have no clear root (the spectral content has no clear fundamental). In both cases, the resulting percept is more complex than that produced by a chord with low roughness and a strong root: in the first case, because of the distracting beating and the difficulty of resolving the individual frequency components; in the second case, because of the difficulty in finding a single representative root pitch. A causal link from complexity to pleasure makes it possible for an artist to play aesthetic games; for example, in music, a greater symmetry may be exposed only gradually such that each segment (e.g., interval, chord, phrase, section) is somewhat complex (asymmetrical) but, when all segments have been played, a more general symmetry or pattern is apparent. There may be other ways in which complexity can be used to communicate higher-level feelings, and there may be other intrinsic nature responses that can be similarly manipulated.

Intrinsic nature processes are due only to some function of the current musical variable. This can be mathematically formalized by the following simple model:

$$h(m_0) = f(m_0)\,, \tag{2.4}$$

which has a single variable m_0.

A very simple example is using log-frequency as a model of pitch height; in other models, such as the affinity model that will be described in Section 4.1.1, the variable may be multidimensional and the function more complex. Purely psychoacoustic models fall into this category, but this category also includes models that use core knowledge (Spelke and Kinzler, 2007) and perceptual principles, like Gestalt, or mathematical procedures, such as similarity measures and pattern detection, which are not psychoacoustic in nature.

2.3.2.5 *Causal interactions*

There are also interactions between these four processes (not shown in Figure 2.1, to avoid a spaghetti effect). For example, a stimulus that is initially perceived as complex will become less so with familiarity (as demonstrated by North and Hargreaves (1995)). In this way, a previously unpleasantly-dissonant tone cluster may become pleasant with sufficient familiarity. This is because familiarity tends to reduce perceived complexity—hence intrinsic nurture processes influence intrinsic nature processes. This can be visualized as a causal path extending from the intrinsic nurture mental process to the intrinsic nature process; or, mathematically, by treating the function f in (2.4) as being parameterized by (2.2). If the intrinsic nature function f is relatively inflexible under the influence of (2.2), it is reasonable to think of the nature process as the more fundamental, or underlying, process.

This relationship also works in reverse, in that a stimulus that evokes a strong innate response is likely more salient and, therefore, probably easier to learn. For example, the pitch associated with the fundamental of a harmonic complex tone is a very salient—probably innate—

perceptual response to such tones; whereas the loudness of the seventh harmonic is not a particularly salient perceptual response. For this reason, we find it easier to learn patterns of pitch than patterns of the seventh harmonic's loudness. This can be visualized as a causal path extending from the intrinsic nature mental process to the intrinsic nurture process; or, mathematically, by weighting the similarity function by the salience implied by (2.4).

Furthermore, salient pitch patterns are more likely to be mentally associated with non-musical events than are non-salient patterns of seventh harmonic loudness, which shows that an intrinsic nature process can also affect an extrinsic nurture process.

In similar ways, each type of process can affect each other type of process, and this illustrates just how difficult it is to cleanly separate out nature from nurture in an experimental investigation. For example, in Chapter 5, I model the affinity of chords and scales by the similarities of the pitches of their partials. However, this presupposes some understanding of what a chord and a scale is, and what 'affinity' means in a musical context. If an individual were raised in a culture that did not use pitch-based music, his or her responses would probably be very different to my participants' and, in an experimental setting, questions such as 'how strongly does this chord produce a feeling of closure or completion?' would probably be completely alien. However, this does not preclude the possibility that, if a pitch-based music were to develop in this culture, it may be more likely to develop in certain directions than others—in directions that are both constrained by, and make use of, underlying innate processes of perception and cognition.

Although causal interactions between the different types of process can occur, as described above, it is important to point out that certain causal paths are unavailable under the definitions given above. For instance, it may seem that previous musical events causally affect the current musical event. For example, it might seem that if a certain musical event is very common in the corpus (i.e., $\sum_t s(m_0, m_t)$ is high), this implies the probability of m_0 is increased and that this would have

a direct impact on a nature process. But the value of m_0, which enters into equations (2.1–2.4) is not its probability of occurring. Rather m_0, in all these models, is taken as a given, which implies its probability of occurring is 'conditioned out'. For this reason, using the definitions I have given above, there can be no causal link between previous musical events and the current musical event.

Although there is no way to completely disentangle the different processes that occur in perception and cognition, the above four-fold categorization is a useful theoretical framework. It is useful because it clarifies the way that different mental processes operate, and the models that are appropriate for them. Notably, it enables us to conceptually differentiate between those aspects of perception and cognition that are in essence innate and universal in that they are, to a meaningful extent, unaffected by typical cultural experiences. It also clarifies how the different processes causally influence each other, and shows how they can operate not just simultaneously but also in concert, with certain stimuli taking on a strong emotional or conceptual charge that is mutually reinforced by all four processes.

In Section 2.3.3, I explore another important causal pathway that occurs over time; this is from the mental processes of composers and performers to the repertoire that will become the previous musical events of a future generation of listeners. In Section 2.3.4, I briefly describe some ways in which the effects of nurture processes can be minimized in an experimental setting.

2.3.2.6 *The mental state*

The result of the four processes, and their interactions, is a mental state, shown at the centre of Figure 2.1. This mental state can be directly observed only by introspection, which is subject to many possible errors and distortions due to the act of observation itself (which disrupts the experience) and preconceptions of what the experience 'should' be (such distortions are discussed at length by Petitmengin and Bitbol (2009)). Furthermore, because it is difficult to communicate to a partic-

ipant a specific categorization of a subjective state, and any such state is frequently a mixture of any number of other states, each of which may have fuzzy boundaries, it is possible the participant will rate *untargeted* mental states that have not been asked for. For example, in the experiment described in Chapter 4, participants were asked to rate how well melodic tones fitted together (their affinity); the data indicates, however, that they were also rating the inherent consonance of the timbres used.

Furthermore, introspection is blind to the processes that lead to this final mental state (see Petitmengin and Bitbol (2009) for a discussion of process blindness): 'people may have little ability to report accurately about their cognitive processes' and even have a 'literal lack of awareness that a process of any kind is occurring until the moment that the result appears' (Nisbett and Wilson, 1977, p. 241). This blindness makes modelling difficult but, then again, we are also blind to processes (natural laws) that underlie the physical world too (if we weren't, modelling would be a trivial exercise!).

Underlying all of these nature and nurture processes are psychoacoustical and cognitive universalities—the peripheral auditory system is sensitive to a limited frequency range, and it exhibits specific acoustical properties that are innate and universal. The central auditory system relies on neural and perceptual processes that are, to a large extent, innate and universal, and there are higher-level cognitive processes that can also be thought of as innate and universal (such as Gestalt principles of perceptual grouping). For example, in any model of musical perception—even nurture processes—it makes sense to model these peripheral mechanisms (e.g., not including ultrasonic frequencies) in the first stages, and take account of general cognitive processes throughout the model. It is, in this sense, that conventional symbolic music notation, such as note names (which characterize a complex tone's fundamental), can provide useful simple variables that already encapsulate psychoacoustic processes. Ultimately, the value of any music variable in the framework models discussed above, is a function of

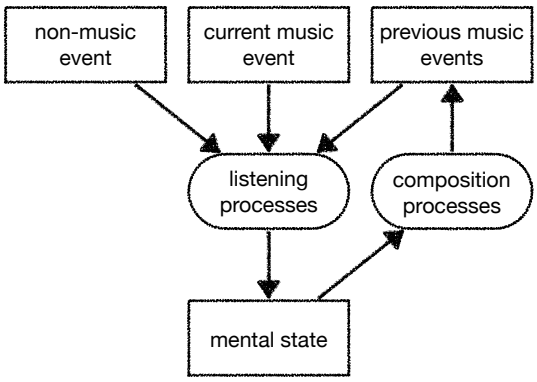


Figure 2.2: Aggregated over people and time, composers create a feedback loop.

an acoustical input variable that can be fully characterized, in the time domain, by its pressure.[18]

Clearly, intertwined with these psychophysical and cognitive universalities, are important learned processes that are non-universal (they are culture- or person-specific) and which strongly affect our perception at fundamental levels. For example, categorical perception determines our experience of the world and is, in part, due to the learning of common patterns found in specific cultural environments.

2.3.3 *Circularity and Explanation*

As illustrated in Figure 2.2, music *composition* can create a feedback loop between mental processes and the musical repertoire.[19] This feedback loop occurs over time and between different individuals and is, therefore, best considered as occurring between mental states and previous musical events that are aggregated over time and numerous individuals. When creating a piece of music, a composer typically imagines or plays a candidate musical event and assesses—by introspection—its impact on his or her mental state. Guided by this, he or she may choose

18 An acoustical signal can be fully characterized in an infinite number of ways, e.g., in the frequency domain, or in the domain of any fractional Fourier transform. The above characterization is, therefore, non-unique.

19 The term *composition* is here used in a general way: it is undertaken not just by *composers*, but also by *consumers* because they choose (by number or influence) which music enters the corpus. The process by which consumers can influence musical development—in a manner akin to natural selection—is nicely demonstrated by the DarwinTunes project (MacCallum et al., 2012).

that event to form part of the composition, or modify it and test it again, or discard it. In this way, the composer draws upon his or her mental processes to produce a final piece of work that is released into the world. Such music is filtered by consumers who 'support' some pieces, but not others, also on the basis of their perceptions. Music that is produced by a composer (by reference to his or her own music perception), and which is supported by consumers (by reference to their music perceptions) is likely to enter the repertoire, or corpus, of music that is commonly heard, and this corpus becomes, in time, the previous musical events that feed into the nurture mental processes of future listeners—some of whom will also be composers. In this way, there is a causal loop (aggregated over time and individuals) from music to mental states (via mental processes of listening) and from mental states back to music (via mental processes of composition), and so on.

To take a plausible concrete example, composers who are sensitive to the innate dissonance produced by psychoacoustical roughness will likely privilege chords with low roughness (i.e., they will use them more often and place them on more salient strong beats and phrase endings). When listeners are familiar with such composers' music they will become especially familiar with low-roughness chords (due to their privileged status), and this will likely increase their perceived consonance. It is worth remembering that thirds and sixths (and, hence, major and minor triads) were considered dissonant prior to the thirteenth century—for typical musical instruments, such intervals are rougher than perfect unisons, octaves, fifths and fourths, but less rough than all other intervals, and their gradual transition to being considered consonant is likely due, in part, to their greater familiarity.[20]

In the above example, the feedback loop is a positive one: innately consonant chords are privileged by composers and hence become perceived as even more consonant due to their familiarity. Such a feedback loop can 'amplify' any underlying process. This process of amplification can be mathematically demonstrated with the simple linear model

20 A different argument, made by Dahlhaus (1990), is that thirds and sixths became perceived as consonant when the underlying tuning system changed from Pythagorean to meantone.

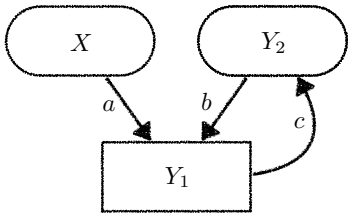


Figure 2.3: A path diagram demonstrating a loop-enhanced effect.

illustrated in the path diagram in Figure 2.3: let the variable X represent the 'smoothness' (lack of roughness) of a chord, Y_1 represent pleasure, and Y_2 the prevalence of $X = x$ in a corpus of music $C = c$, that is, $p_{X|C}(x \mid c)$. Let our model be: $Y_1 = aX + bY_2$ (i.e., pleasure is a linear function of smoothness and prevalence), and $Y_2 = cY_1$ (i.e., prevalence of $X = x$ in the corpus is a linear function of the pleasure created by $X = x$). Error terms have been omitted to simplify the example. All variables are standardized, so their variances equal one and their means zero, hence the parameters a, b, and c are assumed to lie in the interval $[-1, 1]$. In this example, we expect all the parameters to be positive because we expect pleasure to increase with smoothness and prevalence, and prevalence to increase with pleasure. The above two equations permit the following unique solutions for Y_1 and Y_2:

$$Y_1 = \frac{aX}{1 - bc}, \text{ and} \tag{2.5}$$

$$Y_2 = \frac{acX}{1 - bc}. \tag{2.6}$$

In both cases, whenever both b and c are positive, we have a loop-enhanced effect (an amplification of the unlooped effect by $1/(1 - bc)$) of the innate variable smoothness upon both pleasure and the musical corpus.

In reality, the relationship is probably somewhat more complex, because over-familiarity causes boredom. Nevertheless, for any broadly monotonically increasing function between variables in a causal loop, loop-enhanced effects will occur; importantly, this suggests that even a relatively weak nature process may become important to our perception and cognition of music. Similar processes are likely for other

physical and mental variables that are related either by an initial nature process, or by a purely random learned association.

One result of positive feedback loops is that patterns of music come to reflect patterns of our mental processing. This has an important consequence, which is that statistical analyses of a musical corpus can serve as predictively effective models of music perception (e.g., Krumhansl (1990) and Pearce and Wiggins (2006)—see Section 2.3.4 for more details). However, such models are less effective as explanatory tools. Notably, they are unable to provide a reason for why music has taken the specific form it has, or why specific musical events are related to specific mental states in the way they are. This is because such questions are outside the scope of purely learning-based models: for them, it makes no difference if musical structure is founded upon an initial set of associations that was randomly chosen, or whether there are underlying innate processes that make certain musical forms, and certain relationships between musical events and mental states, more viable than others. Indeed, the explanations provided by top-down-only models are circular: music perception is the way it is because music is the way it is, and music is the way it is because music perception is the way it is. As Woolhouse (2009, p.359) puts it:

> In order to elucidate these hypothesized cognitive processes it is not sufficient merely to catalogue their effects (i.e. to describe how tonal music is constituted), or build models that incorporate descriptions of these effects—more important is to uncover the root causes of the effects.

To break into this loop, it is necessary to find some perceptual or cognitive impetus that favours the creation of certain patterns and/or disfavours the creation of others and, possibly, leads to natural associations between specific acoustical events and specific mental phenomena. For this reason, it is important to test bottom-up models of nature processes as part of a search for the underlying processes of human perception, cognition, and emotion. It is only by doing this that we can understand why music is the way it is and, furthermore, build a set of

knowledge that may inform other fields of human enquiry. It is for this reason that my research, in this dissertation, is focused firmly on bottom-up models of innate processes.

In the next section, I explore different types of bottom-up and top-down models, how they may be experimentally tested, and the dangers of relying solely on predictive efficacy when selecting between models.

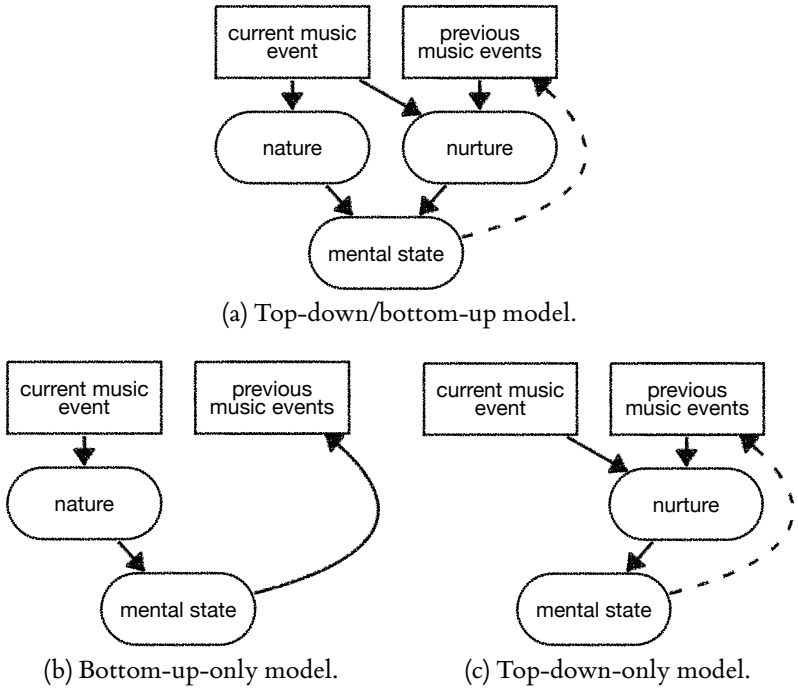
2.3.4 *Bottom-Up and Top-Down Models and Causal Claims*

Numerous experiments have demonstrated a strong correlation between music perception (mental states) and a statistical analysis of a relevant musical corpus. For example, participants' ratings of how well each of the twelve chromatic pitch classes (probe tones) 'fit with' a previously established key is strongly correlated with their prevalence in a corpus of classical music ($r(10) = .89$ in major and $r(10) = .86$ in minor) (Krumhansl, 1990, Ch. 2–3).[21] (In Ch. 5, I provide a detailed description of the probe tone experiment, and I model its data from the bottom-up.) Similarly, participants' ratings of how well differing tones 'continue' a melody are highly correlated with a statistical analysis of a corpus ($r = .85$ with two-tone contexts, and $r = .91$ with a few bars' context) (Pearce and Wiggins, 2006).[22]

Figure 2.4 shows three causal models that can account for correlations such as these (for the sake of clarity, I have omitted non-musical input variables, which determine whether the processes are intrinsic or extrinsic). As in Section 2.3.3, the variables in these models (e.g., 'mental state' and 'previous music events' should be understood as having been aggregated over time and across numerous individuals). In each of these models, there is a direct or indirect causal connection between 'previous music events' and the 'mental state'. The causal connection may be in either direction or both but, to account for the correlation, there must be at least one. In the 'top-down/bottom-up' and

21 The corpus comprised Schubert songs, Mendelssohn arias, Schumann songs, Mozart arias and songs, Hasse cantatas, and Strauss lieder.

22 The corpus comprised Canadian folk songs and ballads, chorale melodies harmonized by J. S. Bach, and German folk songs.



(a) Top-down/bottom-up model.

(b) Bottom-up-only model.

(c) Top-down-only model.

Figure 2.4: Three causal models that can account for correlation between a musical corpus (previous musical events) and mental states (both aggregated over time and across individuals). Physical stimuli are in the top row, mental processes—categorized into 'nature' and 'nurture'—are in the second row, the resulting subjective mental state is in the bottom row. By definition, top-down (nurture) processes are those that receive an arrow from 'previous music events'; bottom-up (nature) processes do not. Note that, as explained in Section 2.3.2.5, a causal path from 'previous music events' to the 'current music event' is not meaningful.

'top-down-only' models, the causal path leading from 'mental state' to 'previous music events' is dashed; this is to indicate that this connection is not necessary for the correlation to occur (there are other causal connections to account for the correlation)—in other words, the 'strength' of the causal effect shown by the dashed arrow can vary from zero to perfect correlation. In the 'bottom-up-only' model, the causal effect of 'mental state' upon 'previous music events' is necessary to ensure correlation (there are no other causal connections to account for the correlation)—the effect size must be non-zero, so this arrow is not dashed. The only way to determine which of these models captures true causal relationships is to perform an experimental intervention: correlation is not sufficient (neither is it necessary). However, it is not uncommon for researchers to be seduced by high correlations into

making unsupported claims about the strengths (or even existence) of causal effects.

For example, due to the greater fit provided by their top-down models, in comparison to a specific bottom-up model, Pearce and Wiggins (2006, p. 378) claim the latter is only 'a formalized approximate description of the mature behaviour of a cognitive system of inductive learning'. A related claim is made by Krumhansl (1990, p. 76):

> Although the acoustic properties associated with consonance may determine to some extent both the way tones are used in music and the quantified tonal hierarchies, the latter two variables are more strongly interdependent. These results point to learning as the dominant mechanism through which the tonal hierarchies are formed.

However, these conclusions are demonstrably unjustifiable on the evidence of correlation alone: in both cases, the experimental evidence is consistent with all of the above models—including the 'bottom-up-only model' depicted in Figure 2.4b.

To demonstrate this, imagine that Figure 2.4's bottom-up-only model of mental states is a true description of reality, but that an experimentally-tested bottom-up model provides an incomplete account of all the relevant bottom-up processes that are actually occurring. For example, suppose the nature process in Figure 2.4b comprises three independent nature processes, but that the experimenters' model captures only one of them. In this situation, the predictive power (e.g., correlation) of the tested bottom-up model may be fairly low because it provides only a partial account of the true processes occurring. However, 'previous music events' naturally captures all three nature processes due to the causal connections leading from 'nature' to 'mental state' to 'previous music events'. This means that, even in the absence of a true top-down nurture process, the statistical analysis of 'previous music events' can have greater predictive power than the experimenters' bottom-up model (because the former is capturing all three, rather than just one, of the nature processes). In this context, therefore,

the greater predictive power of the statistical model is not due to a true top-down process (a causal relationship from 'previous music events' to 'mental states'); rather, it is due to the oppositely-directed process of musical composition (the causal relationship from 'mental states' to 'previous music events').

Clearly, this example shows that a good correlation between a statistical analysis of a corpus and mental states (or a better correlation than a given bottom-up model) cannot be used as evidence for a top-down process occurring in reality. This argument is the other side of the coin to that of Pearce and Wiggins', quoted above; from the evidence of correlation alone, both sides of this coin are equally valid (or invalid).

The principal method to demonstrate causal effects is through experimental intervention—observing the effect of changing some variables while leaving all others fixed (or nullified by randomization). Observing correlation between a model's output and empirical data is simply insufficient. In the context of separating out the nature and nurture processes of music perception, this presents difficulties—to test for nature processes, we need stimuli that have the same degree of familiarity but differ in the nature model's relevant variable.

There are two obvious solutions: to test Western music on participants with no experience of Western music; or to test listeners (who probably do have experience with Western music) with musical stimuli with which they are unfamiliar. In Chapter 4, I use a variation on the latter experimental approach and use a forced choice design with microtonal melodies and a range of unfamiliar spectral tunings (timbres).

2.4 EXISTING MODELS OF TONALITY

There are many verbally defined theories of tonality, but few mathematical or computational models that are able to make precise predictions. In research prior to this dissertation, I identified the main types of theory that have been suggested—from the eighteenth century to the present day (Milne, 2009a). These include *generative-tonic theories*,

which conjecture that the roots of chords seek resolution to pitches of which they are a harmonic, as utilized by Rameau (1726) (as cited in Caplin (1983)), Schenker (1954), and Mathieu (1997); *voice-leading theories*, which propose that tonic pitches are those that are approached by a pitch a semitone away, as suggested by Schenker (1987) and Lerdahl (2001); and a variety of appeals to metaphysics, such as Riemann's theory of tonic functions (as developed by Harrison (1994), Agmon (1995), Quinn (2005), Kelley (2004)).

In this section, I do not wish to retread this ground but I will discuss the two main classes of theory that form part of modern discourse and are also able to make precise predictions that can be empirically tested (i.e., they are mathematical models). Both types of theory are based on the same underlying notion: given a scalic context, the pitch or consonant chord that fits the best (has the highest affinity) is likely to be heard as the most stable and tonic-like. Their difference lies in whether this affinity is modelled as a learned or an innate characteristic. We might, therefore, denote the first class of theories *top-down affinity models*, and the latter class as *bottom-up affinity models*.

The connection between affinity and *tonic-ness* (the extent to which a pitch or chord sounds like a tonic) was first made explicit by Krumhansl. She pointed out the correlation between the fits measured in the probe tone experiment (discussed in Sec. 2.3.4 and Ch. 5) and the 'stability' of the tonic and 'instability' of the chromatic pitch classes—the pitches in the tonic triad are judged to fit the best, the remaining diatonic scale degrees fit the next best, the chromatic degrees fit the worst (Krumhansl, 1990). But a relationship between affinity and tonality is also implicit in tonal theories that place pitches, chords, and keys into regular spatial arrays (lattices) such as the circle-of-fifths or the *Tonnetz*.

The circle-of-fifths is a commonly used representation of the interrelated structure of pitch classes—an example is shown in Figure 2.5. It is interesting to note that if the circle is truncated so as to include only the diatonic pitches, the major or minor triads closest to its centre are the Ionian and Aeolian tonics. That is, in the diatonic scale arranged

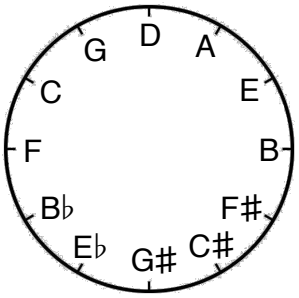


Figure 2.5: The circle-of-fifths.

in fifths F–C–G–D–A–E–B, the most central triads are Cmaj (C–G–E) and Amin (C–A–E).[23] Those are the tonics of the major diatonic scale and the natural (diatonic) minor scale, respectively. Perfect fifths and perfect fourths (which are equivalent intervals when using pitch classes) are typically considered to have the highest consonance and affinity of all intervals (Tenney, 1988). Of the six major or minor triads in the diatonic scale, only the two most central in the chain-of-fifths have no pitch classes at the boundary of the chain, hence only these two have a perfect fifth above and below every one of its pitches. This suggests Cmaj and Amin have the highest affinity with the white-note diatonic scale.

Another commonly used lattice is the *Tonnetz* (German for 'tone-network')—a regular two-dimensional array of pitch classes arranged so as to suggest that all intervals are generated by stacking perfect fifths and major thirds (or perfect fifths and minor thirds)—as illustrated in Figure 2.6.[24] Truncating a *Tonnetz* to contain just the diatonic pitch classes has a similar result to truncating the circle-of-fifths—the most central major or minor triads are the Ionian and Aeolian tonics.[25] In-

23 I use the term *central* to refer to any reasonable measure of the central tendency of a set of points. For example, the *geometric median* minimizes the mean distance between itself and all points; the *centroid* minimizes the mean of squared distances. In Fig. 2.5, the most appropriate measure of distance between any two pitches is the angle between them. The central pitch of the white note diatonic scale is D, but the most central major or minor triads are Cmaj and Amin because their respective centres are closest to D.

24 Such *Tonnetze* have been proposed by numerous theorists, including Euler (1739), von Oettingen (1866), Hostinský (1879), Riemann (1915), Longuet-Higgins (1962), and Balzano (1980).

25 Because this lattice is tiled, its centre is a line rather than a point—as shown by dashed line passing through the Ds. The major or minor triads whose centres are closest to this line are Cmaj and Amin.

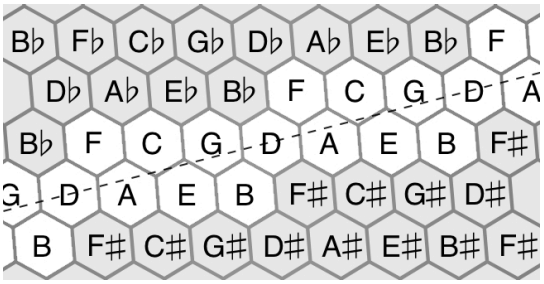


Figure 2.6: A *Tonnetz*.

deed, Balzano (1980) suggests these two chords' centrality accounts for their tonic status. Once again, this suggests that maximizing the number of high affinity perfect fifths and fourths between a triad's pitches and the diatonic scale from which they are taken may be what makes these two chords function as tonics.

Geometrical structures based on the *Tonnetz* have also proved useful for representing higher-order tonal structures. Krumhansl (1990) showed that a multi-dimensional scaling of the perceived fits between keys results in a toroidal structure with an arrangement that closely mirrors a *Tonnetz* that has been wrapped into a cylinder and then into a torus.[26] And Chew (2000) has developed a Spiral Array, which comprises three concentrically arranged cylindrical lattices. The outer lattice is identical to a wrapped *Tonnetz* and contains pitch classes, the middle cylinder is a lattice of chords, and the central cylinder is a lattice of keys. These elements are arranged so each chord is at the *centre of effect* of the pitches it contains, and each key is at the centre of effect of its primary triads (I, IV, and V).[27] The Spiral Array has successfully modelled the distances between pitches, chords and keys, and calculated pitch spellings and the keys of pieces of music played in real time (Chew, 2005, 2006).

26 Wrapping the extended *Tonnetz* into a cylinder ensures that notes separated by the *syntonic comma*—the octave-reduced difference between four perfect fifths and a major third—are mapped to the same spatial location. This makes sense because in common-practice tonal music, intervals separated by this interval are—without exception—treated as identical (in terms of theory, music notation, the note-choices available on instruments, and in performance). Wrapping this cylinder into a torus further removes the spatial distinction between notes that are enharmonic equivalents like A♭ and G♯. Theorists may differ on whether this latter distinction should be kept or removed.

27 The *centre of effect* is the centre of mass where each pitch, or chord, is independently weighted and all weights sum to unity.

In addition to Krumhansl's linking of 'fit' to tonic-ness, the prevalence of *Tonnetz*-based theories—in which the high affinity perfect fifth plays an important role—indicates a widely perceived relationship between affinity and higher-level structures of tonality such as harmonic relationships and the perception of key. As mentioned at the start of this section, I categorize between those models where affinity is assumed to be down to familiarity, and those where it is assumed to be down to an underlying psychoacoustic process.

A notable researcher in the former category is Krumhansl (others are considered in Ch. 5). As discussed earlier, she showed that the fits of all chromatic pitches to a previously established key correlates well with their prevalences in a corpus of classical music (Krumhansl, 1990), and argued this is indicative of a causal relationship from familiarity to perceived fit, and hence to the stability and instability of pitch classes and chords.

A notable researcher in the latter category is Parncutt (others are discussed in Ch. 5). He models affinity by virtual pitch commonality; that is, the number and weights of virtual components in one tone or chord that are the same as those in another tone or chord. *Virtual pitches* are a perceptual phenomenon whereby candidate fundamental pitches are heard in response to a spectrum containing at least some harmonic partials (e.g., Terhardt et al. 1982; Parncutt 1988; Roederer 2008).[28] For example, when the fundamental is removed from a tone with harmonic partials, a virtual pitch corresponding to that fundamental is still heard despite its physical absence. In Parncutt's model of pitch perception, a tone with harmonic partials generates virtual pitch classes at the fundamental and also at subharmonics of the fundamental. Each of these pitch classes is 'quantized' to fall at a single twelve-tone chromatic pitch class. For example, he suggests the notated pitch class C produces virtual pitch classes at F, A♭, D and B♭, because these approximate the the pitch classes of the first nine subharmonics (Parncutt, 1988). This model provides an even better prediction of the probe tone data than

28 *Harmonic partials* are sine waves with frequencies that are integer multiples of a fundamental frequency. Most Western instruments—and the human voice—produce tones with harmonic, or close-to-harmonic, partials.

Krumhansl's familiarity model. And it does so with a plausible psychoacoustic hypothesis.

However, as discussed in greater detail in Chapter 5, there is more to tonality than the data captured in the probe tone experiment. Firstly, the order in which chords are presented is important: for example, compare the progressions Fmaj–Gmaj–Cmaj and Gmaj–Fmaj–Cmaj. In both cases, a listener would typically say the Cmaj is the tonic chord, but that the sense of resolution and closure is stronger in the former progression. And, more generally, the V–I cadence is felt to be the 'strongest' or 'most powerful' progression in tonal music (Piston and Devoto 1987, p. 21; Pratt 1996, p. 9). Temporal asymmetries in tonal perception—such as the above—have been demonstrated in numerous experiments (Brown, 1988; Cuddy and Thompson, 1992; Toiviainen and Krumhansl, 2003), but it is well-acknowledged that models based on an affinity metric (such as *Tonnetz*-based models) cannot account for these temporal asymmetries (Krumhansl, 1990; Woolhouse, 2007). This is because metrics are, by definition, symmetrical, which means the fit of chord x to chord y is the same as the fit of chord y to chord x—the mathematics has no 'arrow' that can represent the flow of time.

Secondly, not just the scale degree but also the chord degree (root, third, or fifth) of pitches seems to change their tonal effect. For example, compare the chord progressions Dmin–Gmaj–Cmaj and Dmin–Emin–Cmaj. A listener would typically report the first progression as providing a more decisive cadence to Cmaj, and that the leading tone B loses much of its desire to resolve to C in the second progression.[29] This is interesting because both chord progressions have the same pitch class content (C, D, E, F, G, A, B), but the tonal effect of B (the seventh degree of that implied scale, referenced from C) seems influenced by whether it is a third (as in the Gmaj in the first progression) or a fifth (as in the Emin of the second progression).

29 The major key progression iii–I is rare (Piston and Devoto, 1987; Huron, 2006), and iii is often used to harmonize a melody where the seventh scale degree descends to the sixth rather than taking its more familiar upwards path to the tonic (Macpherson, 1920).

2.5 MY MODELS OF TONALITY

As I stated in Chapter 1, my research aim is to identify and model the innate processes by which feelings of tension, resolution, stability, and so forth, are induced by successions of pitches and chords, irrespective of their harmonic consonance. In order to fulfill this aim, I develop a novel spectral pitch similarity model of the affinity of any two tones or chords.

Spectral pitch similarity differs from Parncutt's model in two important ways. Firstly, I consider spectral pitches, which correspond directly to frequencies that are in the sounds, rather than virtual pitches, which are generated by the auditory system. This has the advantage of being simpler, because it does not require an additional model for generating the virtual pitches. Secondly, the method of representing these pitches, and measuring their similarities, is quite different.

Spectral pitch similarity is the cosine similarity of two *spectral pitch vectors*, each of which embeds the spectral pitches of a tone or chord. Spectral pitch vectors are mathematically defined in Chapter 3 (specifically Sec. 3.4.1). But, in brief, the index of the vector represents a specific pitch, while the value of each element represents the expected number of partials heard at that pitch. Every frequency that is embedded into a spectral pitch vector is 'smeared' to model the inaccuracies of pitch perception. The width of this smearing is related to the just noticeable frequency difference, which is the smallest frequency difference that can be reliably identified by participants (see App. A, for a more extensive discussion of this parameter). For example, a frequency that corresponds to the MIDI pitch 60, may be represented in a spectral pitch vector by a relatively high value at the index corresponding to the pitch 60, a lower value at the elements corresponding to the pitches 50.99 and 60.01, even lower values at the elements corresponding to 50.98 and 60.02, and so forth. Each spectral pitch vector contains thousands of elements in order to ensure a finely-grained representation of pitch—

in the models discussed in Chapters 4 and 5, I use a granularity of 1 cent (one hundredth of a semitone).

The use of spectral pitch similarity has a number of advantages over Parncutt's approach: (a) it makes no a priori assumption that each frequency component is perceived categorically as a single chromatic pitch (the 'quantization' in my model is to the nearest cent rather than the nearest semitone); (b) it follows a principled (and psychoacoustically justifiable) approach towards modelling the perceived similarity of pitches that are close but non-identical in frequency; (c) it is fully generalizable to musical systems that use tunings different than 12-tone equal temperament—this includes the Pythagorean and meantone tunings that were prevalent in Western music prior to the sixteenth century, contemporary traditions in non-Western music, and experimental contemporary microtonal music.

Furthermore—and as fully explained in Chapter 5—I suggest some additional mechanisms, which build upon spectral pitch similarity, to break the temporal symmetries that are implicit in theories that rely solely upon an affinity metric, and explain why the seventh degree of the major scale loses its perceived need to resolve when it is the fifth of the iii chord. This is achieved by hypothesizing that a resolution, or cadence, is enhanced when a low affinity pitch moves to the root of the best-fitting chord, and that the fit of a pitch class needs to be determined with respect to its local harmonic context as well as to its broader scalic context. In Section 5.2, I show how these latter hypotheses make predictions that accord well with conventional music theory. However, at this stage of the research, there simply isn't any experimental data available to fully test these latter two hypotheses. In the concluding Chapter 6, I discuss the types of data that could provide appropriate tests for these features.

In Section 5.1, I use my basic spectral pitch class similarity model (i.e., without these two additional hypotheses) to model Krumhansl's probe tone data. In comparison with all other existing models, my model produces the most effective predictions (as measured under k-fold cross-

validation) and, being bottom-up, provides an effective explanation for why the probe tone data take the specific form they do.

Chapter 5, in sum, indicates that the spectral pitch similarity model of affinity works well for predicting the tonics and cadences in a variety of scales, and the fits of pitches to a given tonal centre; furthermore, the psychoacoustic basis of the model seems plausible. However, it is still possible that the link between the psychoacoustic mechanism I have proposed and the resulting data is simply coincidental—that I have alighted upon a psychoacoustic explanation that just happens conveniently to fit the data, and that there is, in truth, no real psychoacoustic process occurring.

The purpose of the preceding Chapter 4 is to establish a firm foundation for spectral pitch similarity—to demonstrate that it does actually capture a true psychoacoustic process that accounts for the perceived affinity between tones. To do this I use the method suggested in Section 2.3.4—I use microtonal stimuli that are inherently unfamiliar, and use spectral pitch similarity to model participants' responses to these stimuli. The results indicate that, even in these unfamiliar settings, spectral pitch similarity has a medium-sized (and highly significant) effect on listeners' judgements of affinity. This supports the hypothesis that spectral pitch similarity reflects a real bottom-up process that affects our perception of tonal stimuli. (I additionally show that there is another, unrelated, aspect of the spectral content of tones that affects perceptions of affinity, but the experimental method does not allow for this to be qualified as an innate or learned process).

In order to construct the spectral pitch similarity models utilized in Chapters 4 and 5, it was necessary for me to develop a computationally efficient and principled method for modelling the similarity of pitch collections (such as collections of spectral pitches). This led to the novel family of mathematical methods that are the subject matter of Chapter 3—expectation tensors (multi-way arrays) of differing orders (dimensions), and metrics between them.[30]

30 I developed a spectral pitch 'distance' measure in earlier work (Milne, 2009a), but it was complex to calculate and was not a true metric—it did not fulfill the property

The family of expectation tensors introduced in this chapter are of sufficient generality to extend beyond the uses made of them in Chapters 4 and 5 (both chapters utilize just one type of expectation tensor—absolute monad expectation vectors—which are defined in Sec. 3.4.1). The other types of expectation tensor are also useful and, in Section 3.6, I demonstrate some applications to scale tuning theory and pitch set class theory (these examples are purely illustrative and I do not subject them to experimental testing). I also hint at how expectation tensors might be applied to other (non-pitch) domains, such as rhythm perception and other psychophysical domains.

of the identity of indiscernibles, and I had no proof that it fulfilled the triangle inequality. The desire to develop a simple-to-calculate metric was part of the reason for developing the expectation tensors described in Chapter 3.

3

MODELLING THE SIMILARITY OF PITCH COLLECTIONS

In this chapter, I define a novel family of mathematical methods for determining the similarity of pitch collections. In Sections 3.2–3.5, I present the methods in a general manner; in Section 3.6, I demonstrate some specific high-level applications for the resulting metrics. In Chapters 4 and 5, I develop these methods to provide specific models of more focused empirical data for the perceived fit of microtonal melodies, and the fit and stability of pitches and chords to a variety of musical contexts.

A *pitch collection* may comprise the pitches of tones in a chord, a scale, a tuning, or the virtual and spectral pitches heard in response to complex tones or chords.[1] Modelling the perceived distance (the similarity or dissimilarity) between pairs of pitch collections has a number of important applications in music analysis and composition, in modelling of musical cognition, and in the design of musical tunings. For example, voice-leading distances model the overall distance between two chords as a function of the pitch distance moved by each voice (see Tymoczko (2006) for a survey); musical set theory considers the similarities between the interval (or triad, tetrad, etc.) contents of pitch collections (see Castrén (1994) for a survey); psychoacoustic models of chordal distance (Parncutt, 1989; Milne, 2009a, 2010) treat tones or chords as collections of virtual and spectral pitches (Terhardt et al., 1982; Zwicker and Fastl, 1999) to determine their affinity; tuning theory requires mea-

1 Most sounds—including those produced by musical instruments—comprise numerous frequency components (*partials*) with differing amplitudes and phases. Each of these may produce a corresponding *spectral pitch*. *Virtual pitches* are not physically present in the stimulus but are generated by the auditory system. The most obvious example of a virtual pitch is the missing fundamental—when the fundamental frequency is removed from a stimulus with a high-pass filter, a virtual pitch corresponding to this missing fundamental is typically still heard.

sures that can determine the distance between scale tunings and, notably, the extent to which different scale tunings can approximate privileged tunings of intervals or chords (e.g., just intonation intervals with frequency ratios such as 3/2 and 5/4, or chords with frequency ratios such as 4:5:6:7).

This chapter presents a novel family of embeddings called *expectation tensors* (a tensor is also known as a multi-way array), and associated metrics, that can be applied to the above areas. As discussed in Sections 3.3 and 3.4, expectation tensors model the uncertainties of pitch perception by 'smearing' each pitch over a range of possible values, and the width of the smearing can be related to experimentally determined frequency difference limens (Roederer, 2008). The tensors can embed either absolute or relative pitches (denoted *absolute* and *relative expectation tensors*, respectively): in the latter case, embeddings of pitch collections that differ only by transposition have zero distance; a useful feature that relates similarity to structure. Furthermore, tensors of any order (dimensionality) can be formed, allowing the embeddings to reflect the (absolute or relative) monad (pitch), dyad (interval), triad, tetrad, and so forth, content of the pitch collection.

The distance between expectation tensors of the same order can be determined with any standard metric or similarity measure (such as L_p or cosine). A discussion of how such metrics can be applied and interpreted is found in Section 3.5. In Section 3.6, a number of applications of the metrics are given, and it is shown how distances between different pairs of embeddings (absolute and relative of differing orders) may be combined to produce more informative models of the similarity of pitch collections.

Let me first make some definitions explicit. A *tone* is a sound stimulus that can be characterized by its frequency (e.g., it is playing middle C, which is 261.6 Hz). It may be additionally characterized in a categorical fashion (e.g., it is the third flute from the left, playing middle C). As defined earlier, a *pitch* is the auditory attribute associated with a specific

frequency. For musical purposes, we can say pitch is linearly related to log-frequency, hence we can characterize it by semitones or cents (100ths of a semitone) above a given reference pitch like middle C. I now give a slightly more generalized definition of pitch class. A *pitch-class* is an equivalence class of all pitches that are *periods* apart—a period being a pitch difference over which pitch equivalence is perceived to exist (typically the octave).

Given a tone, the probability of perceiving a specific pitch is denoted the *salience* of that pitch. The pitch perception may be conscious or unconscious; that is, a subject may or may not be aware of it (see Dretske (2006) for a discussion of perception without awareness). For example, the pitch of a partial in a tone may have an impact on the affinity it has with another tone (it communicates information and, hence, is perceived) even when this specific partial is not consciously heard or identified. Throughout the text, when I refer to *perceived*, this does not imply conscious awareness; rather, it implies that the information represented by this pitch is accessible for other mental processes (such as rating affinity). It should also be noted that the values used for the saliences in many of the models I describe in later chapters (where different partials have different saliences) are not directly obtained from empirical research; rather, they are modelled saliences, the parameters of which are optimized to the data being modelled.

Two assumptions are made to simplify the analysis: any given tone can be perceived as having no more than one pitch (or pitch-class), and the perceiving (or not) of a tone does not affect the chance of perceiving another tone.[2] Thus a single note played by an instrument can still be treated as a single perceptual entity (i.e., as a single tone) or as a set of spectral 'tones' corresponding to its partials. *Pitch collections* are treated as multisets—duplication of the same pitch is meaningful because two different tones may induce the same pitch while both remain

2 The latter assumption is reasonable for non-simultaneous tones though, for simultaneous tones that are close in frequency, masking effects will occur which may contradict this assumption. In the models I present in this dissertation, simultaneous tones close in frequency are rare and have a low weighting (they are the higher harmonics in complex tones).

discriminable (e.g., we may perceive a viola and a flute—two tones—both playing the same pitch C4). I use the term *embedding* to refer to a mapping of pitch collections, and their weights, into the discrete multi-dimensional space represented by the tensors. However, contrary to one common usage of this term, I do not mean to imply that these mappings are one-to-one and invertible; in other words, the original pitch and weighting vectors typically cannot be recovered from their embeddings in expectation tensors.

This chapter makes use of tensors and tensor notation. I provide a brief introduction to these in Appendix B. I also provide element-level summations to aid comprehension.

3.1 AN INTRODUCTION TO METRICS AND SIMILARITY

In this chapter, I make extensive use of distance metrics and similarity measures. A *distance metric* is a scalar-valued mathematical function applied to any two members of a set $\mathcal{X}$. Hence it is a mapping $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For the purposes of this dissertation, we are concerned with distances between vectors (or tensors that have been reshaped into vectors), so I write the metrics as scalar-valued functions of vectors. In order to reflect intuitive notions of distance, a metric must fulfill the following properties: non-negativity, identity of indiscernibles, symmetry, and the triangle inequality.

Non-negativity means distance cannot be negative; that is, $d(\mathbf{x}, \mathbf{y}) \geq 0$. *Identity of indiscernibles* means distance is zero, if and only if $\mathbf{x}$ and $\mathbf{y}$ are the same; that is $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$. *Symmetry* means the distance from $\mathbf{x}$ to $\mathbf{y}$ is the same as the distance from $\mathbf{y}$ to $\mathbf{x}$; that is, $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$. The *triangle inequality* means that, given three vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, the distance from $\mathbf{x}$ to $\mathbf{z}$ must be less than or equal to the distance between $\mathbf{x}$ and $\mathbf{y}$ plus the distance from $\mathbf{y}$ to $\mathbf{z}$ (if you imagine $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ as the coordinates of a triangle's vertices, you can see why this is a sensible property for a metric); that is, $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

For any two vectors of the same *dimension* (i.e., they have the same number of elements), a widely used and straightforward metric is the L_p:

$$d_L(\mathbf{x}, \mathbf{y}; p) = \left(\sum_{i=1}^{I} |x[i] - y[i]|^p \right)^{1/p} = \|\mathbf{x} - \mathbf{y}\|_p \qquad \text{for } p \geq 1\,, \tag{3.1}$$

where $|\cdot|$ is the absolute value and $\|\cdot\|_p$ is the p-norm. Different values of p in the L_p metrics correspond to some familiar distances; for example, $p = 1$ is the taxicab distance, $p = 2$ is the Euclidean distance, and $p = \infty$ is the maximum value distance (also known as the Chebyshev distance, it can be written $\max_i |x[i] - y[i]|$).

A *similarity measure* is a closely related function, except it increases when distance decreases and vice versa. Typically, a similarity measure $s(\mathbf{x}, \mathbf{y})$ takes a value of zero when $\mathbf{x}$ and $\mathbf{y}$ are maximally dissimilar, and a value of unity when $\mathbf{x} = \mathbf{y}$. Any distance metric $d(\mathbf{x}, \mathbf{y})$ can be transformed into a similarity measure by taking the function $e^{-d(\mathbf{x},\mathbf{y})}$ (Chen et al., 2009).[3] However, a commonly used similarity measure for vectors is the cosine of the angle between them, which is given by:

$$s_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y})}}\,, \tag{3.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are column vectors and $'$ is the matrix transpose operator, which converts a column vector into a row vector, and vice versa. For vectors, all of whose elements are non-negative, this takes a value between zero (when they orthogonal) and unity (when they are parallel).

Although cosine similarity can be transformed into a true distance metric, called *angular distance*, by

$$d(\mathbf{x}, \mathbf{y}) = \arccos(s_{\cos}(\mathbf{x}, \mathbf{y}))\,, \tag{3.3}$$

3 This is equivalent to Shepard's proposed universal law of generalization, which relates distance in psychological space to perceived similarity (Shepard, 1987).

the simpler *cosine distance*

$$d_{\cos} = 1 - s_{\cos}(\mathbf{x}, \mathbf{y}) \tag{3.4}$$

is a commonly used semi-metric (a *semi-metric* fulfills the first three conditions for a metric, but not the triangle inequality).

Applications of these measures to the expectation tensors is discussed in depth in Section 3.5.

3.2 CATEGORY DOMAIN EMBEDDINGS

Pitch vectors and *pitch class vectors*—both of which are widely used in music theory—are examples of *category domain embeddings* (this is a novel designation, which I introduced in Milne et al. (2011b)). In such embeddings, the values of the elements indicate pitches or pitch classes (usually in semitones), and their *index* (position in the vector) represents a categorical value such as the the type of voice (e.g., bass, tenor, alto, or soprano). For example, the pitch vector $(48, 60, 67, 74)$ can represent a bass part playing the MIDI pitch number 48, the tenor playing 60, the alto 67, and the soprano 74.[4]

As implied by the equations in the previous section, when using standard metrics (such as the L_p and cosine discussed above) between two such vectors, the resulting distances are based only on the pitches in matching positions in the two vectors.[5] For this reason, such pitch metrics are meaningful only when each tone in one pitch collection has a privileged relationship with a unique tone in another pitch collection; for example, when each element (index value) represents a different category such as voice, or scale degree, or even metrical or ordinal po-

4 MIDI pitches are numbered in semitones with a value of 60 corresponding to middle C. So 48 is twelve semitones (one octave) below middle C, and 61 corresponds to the notated pitches C♯ and D♭ a semitone higher than middle C.

5 There are a variety of specialized metrics that do not compare only matching elements, but they are typically insufficiently sensitive (e.g., the Hausdorff distance—see the example provided at the end of this section), or too computationally complex (e.g., cross-bin metrics such as the Earth Mover's Distance (Rubner et al., 2000) which, for vectors with n elements, has a complexity of $O(n^3 \log n)$ in comparison to $O(n)$ for the L_p and cosine (Korte and Vygen, 2007)). In many of my applications, I use vectors containing thousands of elements, so computational simplicity is crucial.

sition in a melody. Furthermore, this can occur only when there are the same number of categories in each tone collection (i.e., both pitch vectors have the same dimension).

Applying metrics to category domain vectors is a well-established technique in music analysis and theory. For example, Chalmers (1990) measures the distances between differently tuned tetrachords using a variety of metrics including L_1, L_2, and L_∞ (so, in this case, he is treating the tetrachord scale-degrees as categories)[6] and Tymoczko (2006) discusses the use of various metrics to measure voice-leading distance (so each voice is in a different category).

To be concrete, a *pitch vector* $\mathbf{x}_\mathrm{p} \in \mathbb{R}^I$ contains elements $x_\mathrm{p}[i]$ indexed by $i = 1, 2, \ldots, I$, where $I \in \mathbb{N}$ is the number of tones. The index i indicates the tone category, and the value of the element $x_\mathrm{p}[i]$ indicates pitch. A typical example is a logarithmic function of frequency

$$x_\mathrm{p}[i] = J \log_b \left(\frac{f[i]}{f_\mathrm{ref}} \right) , \tag{3.5}$$

where $0 < b \in \mathbb{R}$ is the frequency ratio of the period (typically the octave, so $b = 2$), $J \in \mathbb{N}$ determines the number of *pitch units* that make up the period (typically $J = 12$ semitones or $J = 1200$ cents), $f[i] \in \mathbb{R}$ is the frequency of tone i, and $f_\mathrm{ref} \in \mathbb{R}$ is the frequency given a pitch value of zero (typically C_{-1}, which is 69 semitones below concert A, so $f_\mathrm{ref} = 440 \times 2^{-69/12} \approx 8.176$ Hz). With these constants, a four-voice major triad in close position with its root on middle C is $(60, 64, 67, 72)$, which corresponds to the MIDI note numbers for this chord.

A *pitch class vector* or *pc-vector*,

$$x_\mathrm{pc}[i] = x_\mathrm{p}[i] \bmod J , \tag{3.6}$$

is invariant with respect to the period of the pitches since $0 \leq x_\mathrm{pc}[i] \leq J - 1$. This makes it useful for concisely describing periodic pitch collections, such as scales or tunings that repeat every octave. The variable

6 *Tetrachords* are four-note scale fragments, bounded by a perfect fourth. They were used in the construction of Ancient Greek scales, and still form part of some contemporary non-Western traditions.

Table 3.1: These pc-vectors represent several musical scales with $b = 2$ (the frequency ratio of the octave) and $J = 1200$ cents: all pitches from 12-tone equal temperament (12-TET); the 12-TET major scale (Maj-12), all pitches from 10-tone equal temperament (10-TET), and a just intonation major scale (Maj-JI).

12-TET	$(0, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100)$	$\mathbb{R}^{12}$
Maj-12	$(0, 200, 400, 500, 700, 900, 1100)$	$\mathbb{R}^{7}$
10-TET	$(0, 120, 240, 360, 480, 600, 720, 840, 960, 1080)$	$\mathbb{R}^{10}$
Maj-JI	$(0, 204, 386, 498, 702, 884, 1088)$	$\mathbb{R}^{7}$

f_{ref} specifies which pitch class has a value of 0 (in a tonal context, it may be clearest to make it equal to the pitch of the root, or tonic). For example, a major triad may be notated $(0, 4, 7)$ or $(1, 5, 8)$, or more generally as $(x, 4 + x, 7 + x) \bmod J$. Table 3.1 shows some musical scales represented as pc-vectors.

From this point forwards, I will refer to pitch vectors and pc-vectors interchangeably. General statements made about one also apply to the other.

The pc-vector may have an associated *weighting vector*,

$$\mathbf{x}_{\text{w}} \in \mathbb{R}^{I}\,, \tag{3.7}$$

which contains elements $0 \leq x_{\text{w}}[i] \leq 1$. This can be used to represent amplitude, loudness, salience, and so forth. In this dissertation, the weighting vector indicates salience (as defined of this chapter). For example, if four tones sound the pitch classes $(0, 3, 3, 7)$ and have an associated weighting vector $(.9, .6, .6, .9)$, there is probability of $.9$ the pitch of the first tone will be perceived; a probability of $.6$ the pitch of the second tone will be perceived; a probability of $.6$ the pitch of the third tone will be perceived; a probability of $.6$ the pitch of the fourth tone will be perceived.[7]

7 Note that this means there are two ways in which the pitch of 3 may be perceived—from the second tone or the third. As I will explore in greater depth later, this means the overall salience of this pitch is $1 - (1 - .6)(1 - .6) = .84$, and the expected number of tones perceived to be playing this pitch is $.6 + .6 = 1.2$.

However, category domain embeddings, and metrics reliant upon them, are unsuitable when the pitches cannot be uniquely categorized. For example, when modelling the distance between the large sets of spectral pitches perceived in response to complex tones or chords (see Ex. 3.6.1), there is no unique way to reasonably align each spectral pitch of one complex tone or chord with each spectral pitch of another (Sethares et al., 2009) and, even if there were, it is not realistic to expect humans to track the 'movements' of such a multitude of pitches.

A simpler example is provided by the scales in Table 3.1, where the categories are the indices of the scale elements. From a musical perspective, it is clear that some such tunings can be thought of as closer than others. For instance, a piece written in Maj-JI can be played in a subset of 12-TET (such as Maj-12) without undue strain, yet may not be particularly easy to perform when the pitches are translated to a subset of 10-TET. Thus it is desirable to have a metric that allows a statement such as 'Maj-JI is closer to 12-TET than to 10-TET.' (JI is an abbreviation of *just intonation*, TET is an abbreviation of *tone equal temperament*).

When two pc-vectors have the same number of elements, any reasonable metric can be used to describe the distance between them; for example, the distance between Maj-12 and Maj-JI can be easily calculated because they both contain seven pitch classes. However, when two pitch collections have different cardinalities, there is no obvious way to define an effective metric since this would require a direct comparison of elements in $\mathbb{R}^n$ with elements in $\mathbb{R}^m$ for $n \neq m$.[8] One strategy is to identify subsets of the elements of the pitch collections and then try to calculate a distance in this reduced space. For instance, one might attempt to calculate the distance between Maj-JI and 12-TET by first identifying the seven nearest elements of the 12-TET scale, and then calculating the distance in $\mathbb{R}^7$. Besides the obvious problems with identifying corresponding tones in ambiguous situa-

8 In such a case, the Hausdorff metric could be used. This metric is noteworthy because it can be used for sets with differing cardinalities. But, because the distance between any two sets is characterized by the distance between just two points in these sets, it is inadequately sensitive as a model for perceived distance. For example, the Hausdorff distances between C-E-G and D-F♯-A and between C-E-G and C-E-A are identical.

tions, the triangle inequality will fail in such schemes. For example, let pitch collection $\mathbf{y}_{\mathrm{pc}}$ be 12-TET, pitch collection $\mathbf{x}_{\mathrm{pc}}$ be any seven-tone subset drawn from 12-TET (such as the major scale), and pitch collection $\mathbf{z}$ be a different seven note subset of 12-TET. The identification of pitches is clear since $\mathbf{x}_{\mathrm{pc}}$ and $\mathbf{z}_{\mathrm{pc}}$ are subsets of $\mathbf{y}_{\mathrm{pc}}$. The distances $d(\mathbf{x}_{\mathrm{pc}}, \mathbf{y}_{\mathrm{pc}})$ and $d(\mathbf{z}_{\mathrm{pc}}, \mathbf{y}_{\mathrm{pc}})$ are zero under any reasonable metric since $\mathbf{x}_{\mathrm{pc}} \subset \mathbf{y}_{\mathrm{pc}}$ and $\mathbf{z}_{\mathrm{pc}} \subset \mathbf{y}_{\mathrm{pc}}$, yet $d(\mathbf{x}_{\mathrm{pc}}, \mathbf{z}_{\mathrm{pc}})$ is non-zero because the pitch classes in the two scales are not the same. Hence the triangle inequality $d(\mathbf{x}_{\mathrm{pc}}, \mathbf{z}_{\mathrm{pc}}) \leq d(\mathbf{x}_{\mathrm{pc}}, \mathbf{y}_{\mathrm{pc}}) + d(\mathbf{y}_{\mathrm{pc}}, \mathbf{z}_{\mathrm{pc}})$ is violated. Analogous counter-examples can be constructed whenever $n \neq m$.

3.3 PITCH DOMAIN EMBEDDINGS

A way to compare pitch collections with differing numbers of elements is to use a *pitch domain embedding* where the index represents pitch and the value represents the probability of a pitch being perceived, or the expected number of tones perceived at that pitch. Because the cardinality of the pitch domain embedding is independent of the cardinality of the pc-vector it is derived from, such embeddings (and metrics reliant upon them) are able to compare pitch collections with different numbers of tones—such as the spectral and virtual pitches perceived in response to a complex tone or chord, or scales and their tunings. The following examples are shown as transformations of pc-vectors (3.6), but they can also be given in terms of pitch vectors (3.5).

The I elements of a pc-vector $\mathbf{x}_{\mathrm{pc}}$ can be transformed into I indicator functions weighted by the salience vector $\mathbf{x}_{\mathrm{w}}$. This means each pitch class is represented by a row vector, all but one of whose elements is zero. The non-zero element has a value equivalent to the salience of that pitch class, and its index corresponds to its pitch value. These I row vectors are then arranged into a $I \times J$ matrix which, as shown later, allows the saliences of the tones to be individually convolved and

appropriately summed. Formally, the elements of the *pitch class salience matrix* $\mathbf{X}_{\text{pcs}} \in \mathbb{R}^{I \times J}$ are given by

$$x_{\text{pcs}}[i, j] = x_{\text{w}}[i]\, \delta\Big[j - \lfloor x_{\text{pc}}[i] \rceil\Big] \tag{3.8}$$

where $\lfloor z \rceil$ rounds z to the nearest integer and $\delta[z]$ is the Kronecker delta function that is 1 when $z = 0$ and 0 for all $z \neq 0$. Note that the index i ranges over the values 1 to I; however, when using pitch classes, it makes sense for the index j to range over the values 0 to $J - 1$. This is because pitch class values are taken modulo J, hence $j = 0$ is a legal pitch class value, while $j = J$ is not.

Example 3.3.1. Given $J = 12$, $\mathbf{x}_{\text{pc}} = (0, 3, 3, 7)$ (i.e., a close position minor chord with a doubled third), and $\mathbf{x}_{\text{w}} = (1, .6, .6, 1)$, (3.8) gives the pitch class salience matrix

$$\mathbf{X}_{\text{pcs}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.9}$$

Pitch values in the pc-vector are rounded to the nearest pitch unit (whose size is determined by J and b) when embedded in the pitch domain. Using a low value of J (like 12 in Ex. 3.3.1) makes such pitch domain embeddings insensitive to the small changes in tuning that are important when exploring the distances between differently tuned scales, or between collections of virtual and spectral pitches. Naively embedding into a more finely grained pitch domain (such as $J = 1200$) is problematic. For example, under any standard metric, the distance between a tone with a pitch of 400 cents and a tone with a pitch of 401 cents is maximally large (i.e., there is no pair of pitches that will produce a greater distance, see the left side of Fig. 3.1). This is counter to perception since it is likely that two such tones will be perceived as having the same pitch.

The solution is to smooth each spike over a range of pitches to account for perceptual inaccuracies and uncertainties. Indeed, a central

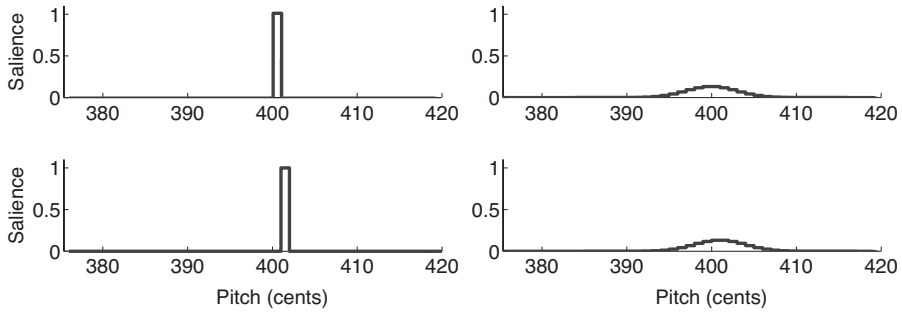


Figure 3.1: Pitch domain embeddings of two tones—one with a pitch of 400 cents, the other with a pitch of 401 cents. On the left, no smoothing is applied, so their distance under any standard metric is maximal; on the right, Gaussian smoothing (standard deviation of 3 cents) is applied, so their distance under any standard metric is small.

tenet of signal detection theory (Green and Swets, 1966) is that a stimulus produces an internal (perceptual) response that may be characterized as consisting of both signal plus noise. The noise component is typically assumed to have a Gaussian distribution, so the internal response to a specific frequency may be modelled as a Gaussian centred on that frequency (Goldstein, 1973). It is this noise component that makes the frequency difference limen greater than zero: when two tones of similar, but non-identical, frequency are played successively, the listener may, incorrectly, perceive them as having the same pitch. The right side of Figure 3.1, for instance, shows the effect of smoothing with a Gaussian kernel with a standard deviation of 3 cents. See Appendix A for a detailed discussion of this parameter and its relationship to the frequency difference limen.

The smoothing is achieved by convolving each row vector in the pitch class salience matrix $\mathbf{X}_{\text{pcs}}$ with a probability mass function. The *pitch class response matrix* $\mathbf{X} \in \mathbb{R}^{I \times J}$ is given by

$$\mathbf{x}[i] = \mathbf{x}_{\text{pcs}}[i] * \mathbf{p}\,, \tag{3.10}$$

where $\mathbf{x}[i]$ is the ith row of $\mathbf{X}$, $\mathbf{x}_{\text{pcs}}[i]$ is the ith row of $\mathbf{X}_{\text{pcs}}$, $\mathbf{p}$ is a discrete probability mass function (i.e., $p_k \geq 0$ and $\sum p_k = 1$ for all k), and $*$ is convolution (circular over the period J when a pc-vector is

used). The result of (3.10) is that each Kronecker delta spike in $\mathbf{X}_{\text{pcs}}$ is smeared by the shape of the probability mass function and scaled so the sum of all its elements is the salience of the tone (as shown in Fig. 3.1). In general, the convolution of the probability distributions of two independent random variables is equivalent to the probability distribution of their sum.[9] This means that when $\mathbf{p}$ is a discrete approximation of a normal distribution, the above convolution gives a discrete approximation of the pitch signal plus normally-distributed noise.

Example 3.3.2. Let the probability mass function be triangular with a full width at half maximum of two semitones; this is substantially less accurate than human pitch perception and a much finer pitch granulation (like cents) would ordinarily be required, but it illustrates the mathematics. Applying this to the pitch class salience matrix of Example 3.3.1 gives the pitch class response matrix

$$\mathbf{X} = \begin{pmatrix} .5 & .25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .25 \\ 0 & 0 & .15 & .3 & .15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .15 & .3 & .15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .25 & .5 & .25 & 0 & 0 & 0 \end{pmatrix}. \quad (3.11)$$

3.4 EXPECTATION TENSORS

The values in the pitch class response matrix represent probabilities; this means it is possible to derive two useful types of embeddings: (a) *expectation tensors* indicate the expected number of tones, ordered pairs of tones, ordered triples of tones, and so forth, that will be perceived as having any given pitch, dyad of pitches, triad of pitches, and so forth; and (b) *salience tensors* indicate the salience of any given pitch, dyad of pitches, triad of pitches, and so forth.

Example 3.3.2 will help to clarify the distinction between expectation and salience: The *expected* number of tones perceived at pitch class 3 is 0.6 (the sum of elements with $j = 3$); this does not mean it is possible to perceive a non-integer number of tones, it means that

9 Let X and Y be independent random variables with probability distributions $f_X(x)$ and $f_Y(x)$. Then, $f_{X+Y} = f_X * f_Y$.

over a large number of experimental trials an average of 0.6 tones will be perceived at pitch class 3 (e.g., given one hundred trials, listeners might perceive two tones at pitch class 3 in nine trials, one tone at pitch 3 in forty-two trials, and perceive no tones at pitch 3 in forty-nine trials). The *salience* (probability of perceiving) a pitch class of 3 is $1 - \Big((1-0)(1-.3)(1-.3)(1-0)\Big) = .51$ so, given one hundred trials, we expect listeners to perceive pitch class 3 a total of fifty-one times (regardless of the number of tones perceived at that pitch). For this dissertation, I consider only expectation tensors.

Expectation tensors may be absolute or relative: *absolute expectation tensors*, denoted $\mathbf{X}_{\mathrm{e}}$, distinguish pitch collections that differ by transposition (e.g., the scales C major and D major), while *relative expectation tensors*, denoted $\widehat{\mathbf{X}}_{\mathrm{e}}$, do not.

Expectation tensors also enable different pitch collections to be compared according to their monad (single pitch), dyad, triad, tetrad, and so forth, content.[10] The formal structure of these different types of embeddings will be made clear in Sections 3.4.1–3.4.4. However, a simple example of why they are useful is provided by considering the major and minor triads $(0, 4, 7)$ and $(0, 3, 7)$ with $J = 12$ and no smoothing. These contain the same set of intervals—a perfect fifth, a major third, and a minor third—and hence they have zero dyadic distance. But, because these intervals are arranged in different ways, they have non-zero triadic distance. Thus the two types of embedding may capture the way major and minor triads are perceived to be simultaneously similar and different. MATLAB routines have been developed to calculate the tensors discussed below; they can be downloaded from `http://www.dynamictonality.com/expectation_tensors_files/`.

10 A *monad* is a set with a single element.

3.4.1 *Monad Expectation Tensors*

The *absolute monad expectation tensor* $\mathbf{X}_{\mathrm{e}}^{(1)} \in \mathbb{R}^J$ is a *vector* that indicates the expected number of tones that will be perceived as corresponding to each possible pitch (class) j.

Example 3.4.1. To take a simple example, let us embed $x_{\mathrm{pc}} = (0, 3, 3, 7)$, $x_{\mathrm{w}} = (1, 1, 1, 1)$, and apply no smoothing—this can be used to represent a four-voice minor triad with a doubled third. These imply that one tone is perceived at pitch 0 (the first tone in the x_{pc} vector), two tones are perceived at pitch 3 (the second and third tones in x_{pc}), one tone is perceived at pitch 7 (the fourth tone in x_{pc}), and no tones are perceived at the other pitches. This can be conveniently represented by the order-1 tensor (vector)

$$\mathbf{X}_{\mathrm{e}}^{(1)} = (1, 0, 0, 2, 0, 0, 0, 1, 0, 0, 0, 0)\,,$$

where the index (which here ranges from 0 to 11) indicates pitch, and its value indicates the number of tones perceived at that pitch.

Absolute monad expectation tensors are useful for comparing the similarity of pitch collections where absolute pitch is meaningful; for example, comparing the spectral or virtual pitches produced by two complex tones or chords in order to model their perceived affinity or fit (see Ex. 3.6.1).

The elements of $\mathbf{X}_{\mathrm{e}}^{(1)}$ are derived from the elements, $x[i, j]$, of the pitch class response matrix by

$$x_{\mathrm{e}}[j] = \sum_{i=1}^{I} x[i, j]\,, \tag{3.12}$$

which is the column sum of the pitch class response matrix $\mathbf{X}$, so

$$\mathbf{X}_{\mathrm{e}}^{(1)} = \mathbf{1}_I'\mathbf{X}\,, \tag{3.13}$$

where $\mathbf{1}_I$ is an I-dimensional column vector of ones, and $'$ is the transpose operator, so $\mathbf{1}_I'$ is a row vector of I ones.

Applied to Example 3.3.2, Equation (3.13) produces $\mathbf{X}_{\mathrm{e}}^{(1)} = (0.5, 0.25, 0.3, 0.6, 0.3, 0, 0.25, 0.5, 0.25, 0, 0, 0.25)$. As shown in Example 3.4.1, when there is no probabilistic smoothing, and every tone has a salience of 1, the monad expectation vector is equivalent to a multiplicity function of the rounded pitch (class) vector; that is, $x_{\mathrm{e}}[j] = \sum_{i=1}^{I} \delta(j - \lfloor x_{\mathrm{pc}}[i] \rceil)$.

A particularly useful application of the absolute monad expectation vector—and one which forms the basis for most of the models in this dissertation—is to embed collections of spectral pitches or pitch classes. For simplicity, I denote the resulting vectors *spectral pitch vectors* or *spectral pitch class vectors*. Almost all musical instruments produce *complex tones*, which comprise numerous *partials* (sine wave components) at differing frequencies. Each of these components can be embedded as a spectral pitch in a spectral pitch vector. This allows for the perceived similarity of any two complex tones (or chords comprising such tones) to be modelled with a similarity measure applied to their spectral pitch vectors. This is the technique I utilize in Chapters 4 and 5 to model experimentally obtained data. To simplify notation in these later chapters, I write spectral pitch vectors as simple vectors (i.e., $\mathbf{x}$) rather than as tensors (i.e., $\mathbf{X}_{\mathrm{e}}^{(1)}$).

In Section 3.6, I also demonstrate some models that use the subsequently described (in Secs. 3.4.2–3.4.4) higher-order tensors. However, I do not formally test these higher-order tensors against empirical data, and they do not form part of the models in Chapters 4–5. This means it is not essential for the reader to fully assimilate their formal properties in order to understand the models that form the core of this dissertation. Having said that, these higher-order tensors are an important independent thread of my research in music cognition, and I hope to test them against new empirical data in future work (see Sec. 6.3).

The *relative monad expectation scalar* $\widehat{\mathbf{X}}_{\mathrm{e}}^{(0)} \in \mathbb{R}$ is a *scalar* that gives the expected overall number of tones that will be perceived (at any pitch).

Example 3.4.2. Given $x_{\mathrm{pc}} = (0, 3, 3, 7)$, $x_{\mathrm{w}} = (1, 1, 1, 1)$, and no smoothing, we perceive a total of four tones. This is naturally represented by the order-0 tensor (scalar)

$$\widehat{\mathbf{X}}_{\mathrm{e}}^{(0)} = 4\,.$$

The relative monad expectation scalar can be calculated by summing $\mathbf{X}_{\mathrm{e}}^{(1)}$ over j or, more straightforwardly, as the sum of the elements of the weighting vector

$$\widehat{\mathbf{X}}_{\mathrm{e}}^{(0)} = \sum_{j=0}^{J-1} x_{\mathrm{e}}[j] = \mathbf{1}_I' \mathbf{X} \mathbf{1}_J = \sum_{i=1}^{I} x_{\mathrm{w}}[i]\,, \tag{3.14}$$

where $\mathbf{1}_J$ is a J-dimensional column vector of ones. Applied to Example 3.3.2, (3.14) gives $\widehat{\mathbf{X}}_{\mathrm{e}}^{(0)} = 3.2$.

3.4.2 *Dyad Expectation Tensors*

The *absolute dyad expectation tensor* $\mathbf{X}_{\mathrm{e}}^{(2)} \in \mathbb{R}^{J \times J}$ is a *matrix* that indicates the expected number of tone pairs that will be perceived as corresponding to any given dyad of absolute pitches.

Example 3.4.3. Let us embed the previously used $x_{\mathrm{pc}} = (0, 3, 3, 7)$, $x_{\mathrm{w}} = (1, 1, 1, 1)$, and apply no smoothing. These imply there are two ordered pairs of tones with pitches 0 and 3 (one pair is the first and second tones in x_{pc}, the other pair is the first and third tones in x_{pc}), there is one ordered pair with pitches 0 and 7 (the first and fourth tones in x_{pc}), there are two ordered pairs of tones with pitches 3 and 3 (one pair is the second and third tones in x_{pc}, the other pair is the third and second tones in x_{pc}), and there are two ordered pairs of tones with the pitches 3 and 7 (one pair is the second and fourth tones in x_{pc}, the other

pair is the third and fourth tones in x_{pc}), and so forth. This can be conveniently represented by the order-2 tensor (matrix)

$$\mathbf{X}_{\mathrm{e}}^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where each entry is identified by two indices, which both range from 0 to 11. The indices specify a pair of pitches (so the top-left element indicates the pitches 0 and 0, the next element to the right indicates the pitches 0 and 1, the element below this indicates the pitches 1 and 1, etc.), while the value of the entry indicates the expected number of ordered tone pairs perceived at that pair of pitches. Specifically, the zeroth row shows there are two ordered tone pairs with pitches 0 and 3 (because $x_{\mathrm{e}}[0,3] = 2$), and there is one tone pair with pitches 0 and 7 (because $x_{\mathrm{e}}[0,7] = 1$). And so forth.

Absolute dyad expectation tensors are useful for comparing the absolute dyadic structures of two pitch collections; for example, to compare scales according to the number of dyads they share—the scales C major and F major contain many common dyads and so have a small distance (.155), the scales C major and F♯ major contain just one common dyad {B, F} and so have a large distance (.782). (These distances are calculated with cosine distance (3.4) and $J = 12$.)

I now describe a method for calculating these tensors. With two tones indexed by $i = 1$ and 2, there are two ordered pairs $(1,2)$ and $(2,1)$. The probability of perceiveing tone 1 as having pitch j and tone 2 as having pitch k is given by $x[1,j]\,x[2,k]$ (these are elements taken from the pitch class response matrix $\mathbf{X}$, which is defined in (3.10)). Similarly, the probability of perceiving tone 2 as having pitch j and tone

1 as having pitch k is given by $x[2, j]\, x[1, k]$. Given two tones, the expected number of ordered tone pairs that will be perceived as having pitches j and k is, therefore, given by $x[1, j]\, x[2, k] + x[2, j]\, x[1, k]$. Similarly, given three tones indexed by $i = 1, 2$, and 3, there are six ordered pairs $(1, 2)$, $(1, 3)$, $(2, 1)$, $(2, 3)$, $(3, 1)$, $(3, 2)$, and the expected number of ordered tone pairs perceived as having pitches j and k is given by the sum of the six probabilities implied by these pairs.

Generalizing for any number of tones, the absolute dyad expectation tensor, $\mathbf{X}_{\mathrm{e}}^{(2)} \in \mathbb{R}^{J \times J}$, contains elements

$$x_{\mathrm{e}}[j, k] = \sum_{\substack{(i_1, i_2) \in \mathcal{I}^2: \\ i_1 \neq i_2}} x[i_1, j]\, x[i_2, k]\,, \tag{3.15}$$

where $\mathcal{I} = \{1, 2, \ldots, I\}$, so $\mathcal{I}^2$ is all ordered pairs from $\mathcal{I}$, and element indices j and k indicate the pitches j and k (note that i ranges from 1 to I, while j and k most conveniently range from 0 to $J - 1$ when using pitch classes). The element value indicates the expected number of ordered pairs of tones perceived as having those pitches. For example, $x_{\mathrm{e}}[0, 7] = 1$ indicates a single pair of pitches at values 0 and 7.

Equation (3.15) requires $O(I^2)$ operations for each element. Using the tensor methods described in Appendix C, this can be expressed directly in terms of $\mathbf{X}$, in a way that requires only $O(I)$ operations per element:

$$\mathbf{X}_{\mathrm{e}}^{(2)} = (\mathbf{1}_I' \mathbf{X}) \otimes (\mathbf{1}_I' \mathbf{X}) - (\mathbf{X}' \mathbf{X})\,, \tag{3.16}$$

where $\otimes$ denotes the *outer product* (also known as the *tensor product*). The outer product is fully explained in Appendix B but, in brief, it multiplies together all possible pairings of elements and applies no summation. For example, the outer product of two vectors $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{y} \in \mathbb{R}^N$ is a matrix $\mathbf{Z} \in \mathbb{R}^{M \times N}$ each of whose elements $z[m, n]$ is the product $x[m]\, y[n]$. For two column vectors, therefore, $\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\, \mathbf{y}'$. As shown in Appendix B, this operation can be generalized to apply to tensors of any order.

The tensor product $(\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X})$ is equivalent to $\sum_{(i_1,i_2)\in\mathcal{I}^2} x[i_1, j]\, x[i_2, k]$, while the matrix product $\mathbf{X}'\mathbf{X}$ is equivalent to $\sum_{i_1=i_2} x[i_1, j]\, x[i_2, k]$, hence subtracting the latter from the former—as in (3.16)—is equivalent to the expression in (3.15). This is explained in greater detail in Appendix C

The *relative dyad expectation tensor* $\widehat{\mathbf{X}}_{\mathrm{e}}^{(1)} \in \mathbb{R}^J$ is a *vector* that indicates the expected number of tone pairs that will be perceived as corresponding to any given dyad of relative pitches (i.e., an interval).

Example 3.4.4. Let us embed the previously used $x_{\mathrm{pc}} = (0, 3, 3, 7)$, $x_{\mathrm{w}} = (1, 1, 1, 1)$, and apply no smoothing. These imply that two ordered pairs of pitches make an interval of size zero (one pair is the second and third tones in x_{pc}, the other is the third and second tones in x_{pc}), there are two ordered pairs of tones making an interval of size 3 (one pair is the first and second tones in x_{pc}, the other is the first and third tones in x_{pc}), there are two ordered pairs of tones making an interval of size 4 (one pair is the second and fourth tones in x_{pc}, the other is the third and fourth tones in x_{pc}), there is one ordered pair of tones making an interval of size 5 (the fourth and first tones in x_{pc}) there is one ordered pair of tones making an interval of size 7 (the first and fourth tones in x_{pc}), there are two ordered pairs of tones making an interval of size 8 (one pair is the fourth and second tones in x_{pc}, the other pair is the fourth and third tones in x_{pc}), there are two ordered pairs of tones making an interval of size 9 (one pair is the second and first tones in x_{pc}, the other pair is the third and first tones in x_{pc}). This can be conveniently represented by the vector

$$\widehat{\mathbf{X}}_{\mathrm{e}}^{(1)} = (2, 0, 0, 2, 2, 1, 0, 1, 2, 2, 0, 0)\,,$$

where the index (which here ranges from 0 to 11) indicates the size of the interval, and its value represents the number of ordered tone pairs perceived as having that interval.

It is useful to note that pitch collections with the same interval content (and weights) have the same relative dyad embedding. For exam-

ple, the pitch vectors $(0, 3, 3, 7)$ and $(0, 4, 4, 7)$ have identical relative dyad embeddings.

Relative dyad expectation tensors are useful for comparing the intervallic structures of two or more pitch collections regardless of transposition. For example, to compare the number of intervals that two pitch collections have in common or to compare different pitch collections by the number, and tuning accuracy, of a specific set of privileged intervals they each contain (for specific applications, see Example 3.6.3, which compares thousands of scale tunings to a set of just intonation intervals).

The relative dyad expectation vector is given by applying row shifts to $\mathbf{X}_{\mathrm{e}}^{(2)}$ so that $k \mapsto k + j$ (circular row shifts so that $k \mapsto k + j \bmod J$, when using pitch classes), and then summing over j, that is,

$$\widehat{x}_{\mathrm{e}}[k] = \sum_{j} x_{\mathrm{e}}[j, k + j]\,, \tag{3.17}$$

where $k + j$ is taken modulo J when pitch class vectors are used. The index k indicates an interval, of size k, above j; for example, $\widehat{x}_{\mathrm{e}}[7] = 1$ indicates one interval of size 7. Assuming the independence of tone saliences, the values are the expected number of ordered tone pairs perceived as having that interval, regardless of transposition.

Equation (3.17) can also be written as autocorrelations:

$$\widehat{\mathbf{X}}_{\mathrm{e}}^{(1)} = \mathbf{1}'\mathbf{X} \star \mathbf{1}'\mathbf{X} - \sum_{i=1}^{I} \mathbf{x}[i] \star \mathbf{x}[i]\,, \tag{3.18}$$

where $\star$ denotes cross-correlation and $\mathbf{x}[i]$ is the ith row of the pitch class response matrix $\mathbf{X}$. This is the autocorrelation of the column sum of the pitch class response matrix $\mathbf{X}$ minus the autocorrelations of each of its rows.[11] The advantage of this form is that we can make use of the autocorrelation theorem to efficiently calculate this vector using fast

11 In this respect, it bears an interesting relationship to Lewin's interval function of a pitch class set with itself (in Lewin's notation, $\mathrm{IFUNC}(X, X)$), but with an additional correction—the subtracted second term—that stops each tone from being compared to itself (Lewin, 1959, 2001).

Fourier transforms (FFTs).[12] Unfortunately, there is no obvious way to generalize this simplification for the higher-order relative expectation tensors discussed later.

As shown in Example 3.4.4, when there is no probabilistic smoothing applied, and the salience of every tone is 1, the relative dyad expectation vector simply gives the multiplicity of ordered pairs of tones that correspond to any possible interval size. When there are no tones with the same pitch class (this is always the case, by definition, when using a pitch class set rather than a multiset), the zeroth element of the unsmoothed relative dyad expectation vector always has a value of 0. Because the values of all its elements are symmetrical about the zeroth element, no information is lost by choosing the subset $\{\widehat{x}_{\mathrm{e}}[k] : 1 \leq k \leq \lfloor J/2 \rfloor\}$ and, when J is an even number, dividing the last element by two (otherwise it is double-counted). When $J = 12$, this subset is identical to the 6-element *interval-class vector* of atonal music theory (Forte, 1973). The relative dyad expectation tensor can, therefore, be thought of as a generalization of a standard interval vector that can deal meaningfully with doubled pitches and the uncertainties of pitch perception.

3.4.3 *Triad Expectation Tensors*

The *absolute triad expectation tensor* $\mathbf{X}_{\mathrm{e}}^{(3)} \in \mathbb{R}^{J \times J \times J}$ indicates the expected number of ordered tone triples that will be perceived as corresponding to any given triad of absolute pitches.

Example 3.4.5. Let us embed the previously used $x_{\mathrm{pc}} = (0, 3, 3, 7)$, $x_{\mathrm{w}} = (1, 1, 1, 1)$, and apply no smoothing. These imply that there are two ordered triples of tones with pitches 0, 3, and 3 (one triple is the first, second, and third tones in x_{pc}, the other is the first, third, and second tones in x_{pc}). There are two ordered triples of tones with pitches 0, 3, and 7 (one triple is the first, second, and fourth tones in x_{pc}, the

12 The autocorrelation theorem states that $\mathbf{f} \star \mathbf{f} = \mathcal{F}\left(\left|\mathcal{F}(\mathbf{f})\right|^2\right)$, where $\star$ is circular cross-correlation (so $\mathbf{f} \star \mathbf{f}$ is the autocorrelation of $\mathbf{f}$), and $\mathcal{F}$ denotes the Fourier transform.

other triple is the first, third, and fourth tones in x_{pc}). There are two ordered triples with pitches 3, 3, and 7 (one triple is the second, third, and fourth tones in x_{pc}, the other is the third, second, and fourth tones in x_{pc}). There are no ordered triples at any other triad of pitches. This can be conveniently represented by an order-3 tensor (a three-way array, or cube of numbers). In such a tensor, each entry is identified by three indices each of which ranges from 0 to 11. The indices indicate a triad of pitches, while the value of that entry indicates the expected number of ordered tone triples perceived at that triad of pitches. So, for this example, the tensor is all zeros except for entries with index values that are a permutation of $(0, 3, 3)$, $(0, 3, 7)$, and $(3, 3, 7)$ which have a value of 2 (i.e., $x_{\mathrm{e}}[0, 3, 3] = x_{\mathrm{e}}[3, 0, 3] = x_{\mathrm{e}}[3, 3, 0] = x_{\mathrm{e}}[0, 3, 7] = x_{\mathrm{e}}[0, 7, 3] = x_{\mathrm{e}}[3, 0, 7] = x_{\mathrm{e}}[3, 7, 0] = x_{\mathrm{e}}[7, 0, 3] = x_{\mathrm{e}}[7, 3, 0] = x_{\mathrm{e}}[3, 3, 7] = x_{\mathrm{e}}[3, 7, 3] = x_{\mathrm{e}}[7, 3, 3] = 2$).

Absolute triad expectation tensors are useful for comparing the absolute triadic structures of two pitch collections; for example, to compare two scales according to the number of triads they share—the scales C major and F major have many triads in common (e.g., {C, E, G}, {C, D, E}, and {D, F, G} are found in both scales) and so have a small distance (.170), the scales C major and F♯ major have no triads in common—they share only two notes {B, F}—and so have the maximal distance of 1. (These distances are calculated with the cosine distance (3.4) with $J = 12$.)

I now describe a method for calculating these tensors. Given three tones indexed by 1, 2, and 3, there are six ordered triples $(1, 2, 3)$, $(2, 1, 3)$, $(2, 3, 1)$, $(1, 3, 2)$, $(3, 1, 2)$, $(3, 2, 1)$; the probabilities of perceiving each triple as having pitches j, k and ℓ, respectively, are $x[1, j]\, x[2, k]\, x[3, \ell]$, $x[2, j]\, x[1, k]\, x[3, \ell]$, $x[2, j]\, x[3, k]\, x[1, \ell]$, $x[1, j]\, x[3, k]\, x[2, \ell]$, $x[3, j]\, x[1, k]\, x[2, \ell]$, and $x[3, j]\, x[2, k]\, x[1, \ell]$. Given three tones, the expected number of ordered tone triples perceived as having pitches j, k, ℓ is given by the sum of the above probabilities.

Generalizing for any number of tones, the absolute triad expectation tensor, $\mathbf{X}_{\mathrm{e}}^{(3)} \in \mathbb{R}^{J \times J \times J}$ contains elements

$$x_{\mathrm{e}}[j,k,\ell] = \sum_{\substack{(i_1,i_2,i_3)\in\mathcal{I}^3:\\ i_1\neq i_2, i_1\neq i_3, i_2\neq i_3}} x[i_1,j]\, x[i_2,k]\, x[i_3,\ell]\,, \tag{3.19}$$

where $\mathcal{I} = \{1, 2, \ldots, I\}$, and $\mathcal{I}^3$ is all ordered triples of elements from $\mathcal{I}$, and j, k, and ℓ indicate the pitch (classes) j, k, and ℓ (note that i ranges from 1 to I, while j, k, and ℓ most conveniently range from 0 to $J-1$ when using pitch classes). Assuming the independence of tone saliences, element value indicates the expected number of ordered triples of tones perceived as having those three pitches. For example, $x_{\mathrm{e}}[0,4,7] = 1$ indicates a single triad of pitches at values 0, 4, and 7.

Equation (3.19) requires $O(I^3)$ operations for each element, but can be simplified to $O(I)$ by using the tensor methods of Appendix C:

$$\begin{aligned}
\mathbf{X}_{\mathrm{e}}^{(3)} = {} & (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X}) \Big)_{\langle 1,2,3\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X}) \Big)_{\langle 2,1,3\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X}) \Big)_{\langle 3,1,2\rangle} \\
& + 2\,(\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\,,
\end{aligned} \tag{3.20}$$

where $\odot$ denotes the Khatri-Rao product, and the angle bracketed subscripts denote mode permutations, both of which are explained in full in Appendix B. This equation is a higher-order generalization of (3.16) in that the first line calculates the summation in (3.19) but over all terms (i.e., $(i_1, i_2, i_3) \in \mathcal{I}^3$). The next three lines remove summations over any terms where $(i_1 = i_2)$, $(i_1 = i_3)$, and $(i_2 = i_3)$. But this removes the term $i_1 = i_2 = i_3$ too many (three) times, so two of them are put back by the final line. The process used here is similar to the *inclusion-exclusion principle* used in combinatorics to count the numbers of ele-

ments in intersecting sets. The method is explained in greater detail in Appendix C.

The *relative triad expectation tensor* $\widehat{\mathbf{X}}_{\mathrm{e}}^{(2)} \in \mathbb{R}^{J\times J}$ is a *matrix* that indicates the expected number of ordered tone triples perceived at any possible relative triad (the latter is characterized by a pair of single-reference intervals). By *single-reference*, I mean that both intervals are measured with respect to the same pitch; for instance, a three-voice minor triad can be characterized by the interval pair 3 and 7 because it has intervals of those sizes with respect to pitch 0 (i.e., $3-0$ and $7-0$). The same triad can also be represented by the interval pair 4 and 9 because it has intervals of those sizes with pitch 4 (i.e., $7-3$ and $0-3 \bmod 12$), or by the interval pair 5 and 8 because it has intervals of those sizes with pitch 7 (i.e., $0-7 \bmod 12$ and $3-7 \bmod 12$).

Example 3.4.6. Let us embed the previously used $x_{\mathrm{pc}} = (0,3,3,7)$, $x_{\mathrm{w}} = (1,1,1,1)$, and apply no smoothing. These imply there are two ordered tone triples with the single-reference interval pair 0 and 4 (one triple is the second, third, and fourth elements of x_{pc}, the other is the third, second, and fourth elements of x_{pc}), there are two ordered tone triples with the interval pair 0 and 9 (one triple is the second, third, and first elements of x_{pc}, the other is the third, second, and first elements of x_{pc}), there are two ordered tone triples with the interval pair 3 and 3 (one triple is the first, second, and third elements of x_{pc}, the other is the third, third, and second elements of x_{pc}), there are two ordered tone triples with the interval pair 3 and 7 (one triple is the first, second, and fourth elements of x_{pc}, the other is the first, third, and fourth elements of x_{pc}), there are two ordered tone triples with the interval pair 4 and 9 (one triple is the second, fourth, and first elements of x_{pc}, the other is the third, fourth, and first elements of x_{pc}), there are two ordered tone triples with the interval pair 5 and 9 (one triple is the fourth, first, and second elements of x_{pc}, the other is the fourth, first, and third

elements of x_{pc}). This can be conveniently represented by the order-2 tensor (matrix)

$$\widehat{\mathbf{X}}_{\mathrm{e}}^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

where each entry is identified by two indices, which both range from 0 to 11. The indices specify a pair of single-reference intervals (so the top-left element indicates the intervals 0 and 0, the next element to the right indicates the intervals 0 and 1, the element below this indicates the intervals 1 and 1, etc.), while the value of the entry indicates the expected number of ordered tone triples perceived as having that pair of intervals. Specifically, the zeroth row shows there are two ordered tone triples with single-reference intervals of sizes 0 and 4 (because $\widehat{x}_{\mathrm{e}}[0, 4] = 2$), and there are two tone triples with single-reference intervals of sizes 0 and 9 (because $\widehat{x}_{\mathrm{e}}[0, 9] = 2$); the third row shows there are two ordered tone triples with single-reference intervals of sizes 3 and 3 (because $\widehat{x}_{\mathrm{e}}[3, 3] = 2$), and there are two triples with single-reference intervals of sizes 3 and 7 (because $\widehat{x}_{\mathrm{e}}[3, 7] = 2$). And so forth.

It is interesting to note that different pitch collections with the same interval content (and weights) may have different relative triad embeddings. For example, the pitch vectors $(0, 3, 3, 7)$ and $(0, 4, 4, 7)$ have non-identical relative triad embeddings.

Relative triad expectation tensors are useful for comparing the triadic structures of two or more pitch collections, regardless of transposition. For example, to compare the number of triad types two pitch collections have in common; or to compare pitch collections by the number, and tuning accuracy, of a specific set of privileged triads they

each contain (for a specific application, see Ex. 3.6.3, which compares thousands of scale tunings against a just intonation triad).

The relative triad expectation matrix is given by applying mode shifts to $\mathbf{X}_{\mathrm{e}}^{(3)}$ so that $k \mapsto k + j$ and $\ell \mapsto \ell + j$ (circular mode shifts so that $k \mapsto k + j \bmod J$ and $\ell \mapsto \ell + j \bmod J$, when embedding pitch classes), and then summing over j:

$$\widehat{x}_{\mathrm{e}}[k, \ell] = \sum_{j} x_{\mathrm{e}}[j, k + j, \ell + j] \,, \tag{3.21}$$

where $k + j$ and $\ell + j$ are taken modulo J when used with pitch class vectors. Element indices k and ℓ indicate two intervals, of sizes k and ℓ, above j (which together make a triad). Assuming independence of tone saliences, the element values are the expected number of ordered tone triples perceived as corresponding to that relative triad. For example, $\widehat{x}_{\mathrm{e}}[4, 7] = 1$ indicates one ordered tone triple containing single-reference intervals of sizes 4 and 7.

In the same way the relative dyad expectation vector can be considered to be a psychoacoustically informed generalization of the interval class vector, the relative triad expectation matrix can be thought of as a generalization of the *trichord-class vector* (also known as a 3-class vector), which is used in musical set theory (Lewin, 1987; Castrén, 1994; Buchler, 2001; Kuusi, 2001).[13] The tensor representation also has the advantage that the indexing of its elements is directly related to the pitch structure, rather than just following Forte's set class designations (1973).

3.4.4 *R-ad Expectation Tensors*

The definitions and techniques of the previous sections can be generalized to R-ads of any order (so long as $R \leq I$).

13 It is also equivalent to Lewin's $\mathrm{EMB}(\mathrm{CANON}, /\mathrm{X}/, \mathrm{Y})$ function, taken over all trichords in the set class /X/, where the canonical group CANON is transposition operations (not inversions) (Lewin, 1987, p. 106).

An *absolute R-ad expectation tensor*, $\mathbf{X}_{\mathrm{e}}^{(R)} \in \mathbb{R}^{J^R}$ indicates the expected number of ordered tone R-tuples that will be perceived as corresponding to any given R-ad of absolute pitches. It contains elements

$$x_{\mathrm{e}}[j_1, j_2, \ldots, j_R] = \sum_{\substack{(i_1,\ldots,i_R)\in\mathcal{I}^R:\\ i_n\neq i_p}} \prod_{r=1}^{R} x[i_r, j_r]\,, \tag{3.22}$$

where $\mathcal{I} = \{1, 2, \ldots, I\}$. Element indices j_1, j_2, $\ldots$, j_R indicate the pitches j_1, j_2, $\ldots$, j_R (note that i ranges from 1 to I, while j_1, j_2, $\ldots j_R$ most conveniently range from 0 to $J-1$ when using pitch classes). Assuming the independence of tone saliences, element value indicates the expected number of ordered R-tuples of tones perceived as having those R pitches. For example, $x_{\mathrm{e}}[0, 4, 7, \ldots, 11] = 1$ indicates a single ordered R-tuple of tones with pitches 0, 4, 7, …, and 11.

As explained in Appendix C, this can also be expressed directly in tensor notation:

$$\mathbf{X}_{\mathrm{e}}^{(R)} = \left(\left(\mathbf{1}_{J^R} \otimes \mathbf{E}_{I^R}\right) \circ \mathbf{X}^{\otimes R}_{\langle R+1,1,R+2,2,\ldots,\ldots,R+R,R\rangle}\right) \overset{R}{\bullet} \mathbf{1}_{I^R}\,. \tag{3.23}$$

Equations (3.22) and (3.23) are symbolically concise, but cumbersome to calculate since each element of $\mathbf{X}_{\mathrm{e}}^{(R)}$ requires $O(I^R)$ operations. Fortunately, this can be reduced to $O(I)$ by breaking (3.23) into subspaces which are then added and subtracted in a manner analogous to that shown in (3.16) and (3.20). This process is fully explained in Appendix C. As shown in the MATLAB routines at `http://www.dynamictonality.com/expectation_tensors_files/`, the computational complexity can be further reduced by exploiting the sparsity of the tensors to calculate only non-zero values; furthermore, due to their construction, the tensors are invariant with respect to any transposition of their indices, so only non-duplicated elements need to be calculated. To minimize memory requirements, the tensors can be stored in a sparse format.

A *relative R-ad expectation tensor* indicates the expected number of ordered tone R-tuples perceived at any possible relative R-ad (the latter characterized by $R-1$ single-reference intervals). It is invariant with respect to transposition of the pitch collection. The absolute R-ad expectation tensors are transformed into relative R-ad expectation tensors by shifting modes $2, 3, \ldots, R$ of $\mathbf{X}_{\mathrm{e}}^{(R)}$ so that $j_r \mapsto j_r + j_1$ (circularly shifting modes so that $j_r \mapsto j_r + j_1 \bmod J$, when embedding pitch classes), and then summing over j_1. This creates an order-$(R-1)$ *relative R-ad expectation tensor* with elements

$$\widehat{x}_{\mathrm{e}}[j_2, j_3, \ldots, j_R] = \sum_{j_1} x_{\mathrm{e}}[j_1, j_2 + j_1, \ldots, j_R + j_1] \in \mathbb{R}^{J^{R-1}} . \quad (3.24)$$

Element indices $j_2, \ldots, j_R$ indicate a set of $R-1$ intervals above j_1 (which together make an R-ad); assuming the independence of tone saliences, element value indicates the expected number of ordered R-tuples of tones that are perceived as corresponding to that relative R-ad (set of $R-1$ single-reference intervals). For example, $\widehat{x}_{\mathrm{e}}[j_2, j_3, \ldots, j_R] = 1$ indicates one R-tuple of tones containing single-reference intervals of sizes j_2, j_2, …, j_R.

As before, the relative R-ad expectation tensors can be viewed as generalizations of subset-class vectors of cardinality R (also called n-class vectors, where $n = R$). In comparison to these vectors, the tensors have the advantage of a principled system of indexing, as well as meaningfully accounting for duplicated pitches and the uncertainties of pitch perception.

3.5 METRICS

The focus of the previous section has been on different ways of embedding a single collection of pitches into a tensor. In this section, I discuss methods to measure the distance between, or similarity of, any two such tensors. In particular I discuss the two common distances in-

troduced in Section 3.1—the L_p and the cosine—which are used in the applications of Section 3.6.

It is reasonable to model the perceived pitch distance between any two tones with their absolute pitch difference (e.g., the pitch distance between tones with pitch values of 64 and 60 semitones is 4 semitones). The L_p-metrics are calculated from absolute differences so they provide a natural choice for calculating the overall distance between pairs of category domain pitch vectors. When there are I different tones in each vector, there are I different pitch differences; the value of p determines how these are totalled (e.g., $p = 1$ gives the taxicab measure, which simply adds the distances moved by the different voices; $p = 2$ gives the Euclidean measure; $p = \infty$ gives the maximum value distance, which is largest distance moved by any voice). As discussed in Section 3.2, the use of such metrics is a well-established procedure.

The metrics may be based on the intervals between pairs of pitch, or pitch class, vectors in $\mathbb{R}^I$:

$$d_{\mathrm{w}}(\mathbf{x}_{\mathrm{pc}}, \mathbf{y}_{\mathrm{pc}}; p) = \left(\sum_{i=1}^{I} w[i] \left| x_{\mathrm{pc}}[i] - y_{\mathrm{pc}}[i] \right|^p \right)^{1/p}, \tag{3.25}$$

where the weights $w[i]$ may be sensibly chosen to be the product of the saliences $w[i] = x_{\mathrm{w}}[i]\, y_{\mathrm{w}}[i]$ from (3.7) (Parncutt, 1989). The metrics may also treat the unordered pitch class intervals:[14]

$$d_{\mathrm{c}}(\mathbf{x}_{\mathrm{pc}}, \mathbf{y}_{\mathrm{pc}}; p) = \left(\sum_{i=1}^{I} w[i] \min_{k \in \mathbb{Z}} \left| x_{\mathrm{pc}}[i] - y_{\mathrm{pc}}[i] - kJ \right|^p \right)^{1/p}. \tag{3.26}$$

Equation (3.25) provides a measure of pitch height distance while (3.26) provides a measure of pitch class (or chroma) distance.

To calculate the distance between two pitch domain expectation tensors $\mathbf{X}_{\mathrm{e}}^{(R)}$ and $\mathbf{Y}_{\mathrm{e}}^{(R)} \in \mathbb{R}^{\overbrace{J \times J \times \cdots \times J}^{R}}$, the L_p and cosine distances can

14 *Unordered pitch class intervals*, also known as *interval classes* or *undirected intervals*, are commonly used in musical set theory. They give the same value to any given interval and its inversion; for example, both a perfect fifth and a perfect fourth are represented by the value of 5 semitones.

be applied in an entrywise fashion. The simplest way to write this is to reshape the tensors into column vectors $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^{J^R}$, which may be applied in (3.1) and (3.2).

The cosine similarity of two vectors is equivalent to their uncentred correlation, and the use of such metrics is an established procedure in music theory and cognition (Krumhansl, 1990; Scott and Isaacson, 1998; Rogers, 1999).[15] For expectation tensors, the meaning of the cosine similarity is easier to discern (and is a more obvious choice) than that of the L_p-metrics: It gives a normalized value for the expected number of ways in which each different R-ad in one pitch collection can be matched to a corresponding R-ad in another pitch collection. For example, consider the absolute triad expectation tensors for the scales C major and D major, where each tone has a salience of 1 and no probabilistic smoothing is applied. The numerator of the division counts the number of triad matches: both contain the triad {G, A, B}, which gives a count of 1; both contain the triad {A, C, E}, which increases the count to 2; both contain the triad {A, B, E}, which gives a cumulative total of 3; and so on, for all possible triads. The denominator of the division then normalizes the value to the interval $[0, 1]$. Similarly, for a relative triad expectation tensor, both C major and D major contain three root-position major triads each, so there are a total of 9 ways they can be matched; both contain one root-position diminished triad each, so there is 1 way they can be matched, making a cumulative total of 10; and so on, for all possible relative triads. The denominator of the division again normalizes.

The final choice of distance or similarity measure can be made a priori (guided by theory, as above) or post-hoc (as a free parameter chosen to fit empirical data).

15 The Pearson correlation of two vectors is equivalent to the cosine similarity of the two vectors with their means subtracted. That is, $r(\mathbf{x}, \mathbf{y}) = s_{\cos}(\mathbf{x} - \overline{\mathbf{x}}, \mathbf{y} - \overline{\mathbf{y}})$, where $\overline{\mathbf{x}}$ is a vector all of whose values are the mean of $\mathbf{x}$ (and the same for $\overline{\mathbf{y}}$ and $\mathbf{y}$).

3.6 APPLICATIONS

In this section, I introduce some ways in which expectation tensors can be used to model various aspects of music perception. In Section 3.6.1, I briefly discuss spectral pitch vectors (embeddings, in absolute monad expectation tensors, of the partials of complex tones) and how spectral pitch distance (the distance between spectral pitch vectors) can be used to model the affinity of chords, and how this measure differs from the commonly used voice-leading distance (the distance between pitch-vectors or pc-vectors). The spectral pitch vectors, introduced here, form the basis of the models I develop and experimentally test, in Chapters 4 and 5.

I also demonstrate some alternative applications for the other types of expectation tensor—the relative expectation tensors and higher-order tensors. In Section 3.6.2, I show how they can be used to model the extent to which different scale tunings are heard to approximate a set of privileged intervals such as just intonation. This is useful for the design of microtonal scale systems. In Section 3.6.3, I also show how they can be used to replicate and generalize aspects of pitch set theory, which is a technique to analyze the perceived distances between pitch class sets in atonal music theory. Unlike the spectral pitch similarity models, these latter two types of model are not developed or empirically tested in the remainder of the dissertation. For that reason, Sections 3.6.2 and 3.6.3 may be regarded as useful but non-essential reading.

The MATLAB routines used to calculate these examples can be downloaded from `http://www.dynamictonality.com/expectation_tensors_files/`.

3.6.1 *Tonal Distances*

Although I will cover spectral pitch vectors in greater detail in subsequent chapters, it is useful at this stage to demonstrate how spectral

pitch distance differs from the more familiar voice-leading distance, which is a commonly used measure in music theory.

For modelling the distance between chords, the most common approach is to measure their *voice-leading distance*. This is the L_p distance (usually with $p = 1$ or $p = 2$) between category domain embeddings of the chords. For instance, the distance between the chord Cmaj and the chord Dmin can be modelled as $\|(0, 4, 7) - (2, 5, 9)\|_p$ (the p-norm of their difference, as in (3.1)). This is clearly an intuitive and obvious way to measure perceived distance.

An alternative approach is to consider the distance between two chords when the partials of each chord are embedded into a spectral pitch (class) vector—that is, to calculate their *spectral pitch distance*. Any complex tone or chord produces a large number of spectral and virtual pitch responses (Terhardt et al., 1982; Zwicker and Fastl, 1999), which suggests that the distances between collections of spectral or virtual pitches may provide an effective model for the perceived affinity of tones or chords (Parncutt, 1989; Milne, 2009a).[16] For most musical stimuli, there are so many spectral or virtual pitches it is unlikely they can be mentally categorized. The appropriate embedding is, therefore, in the pitch domain not a categorical domain.

Example 3.6.1. *Voice-leading distance and spectral pitch distance.*

For a selection of triads, Figure 3.2 illustrates the difference between voice-leading distance and spectral pitch distance. Each block is a triad that contains a perfect fifth. The pitch of both tones in this fifth increase, by semitones, from left to right. For example, if the pitches of the fifth in the central column are C–G, the next column to the right is C♯–G♯, and the column after that has the fifth D–A. The pitch of the 'third' increases, by semitones from bottom to top. For example, if the pitch of the 'third' is E in the central row, the next row above has a 'third' of F, and the row above that has a 'third' of F♯. This means the major and minor triads (the only triads deemed consonant in common-

16 *Virtual pitches* are generated by the auditory system in response to complex tones. The most common example is the pitch heard to correspond to a missing fundamental. Virtual pitches are discussed in greater detail in Chapters 4 and 5.

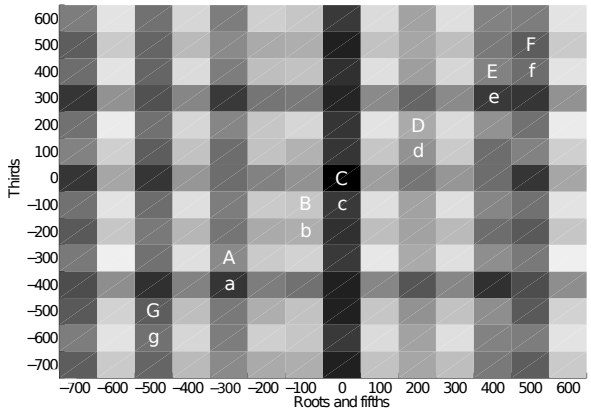


Figure 3.2: Spectral pitch distances of a Cmaj reference triad and all possible 12-TET triads that contain a perfect fifth. (Spectral pitch distance is calculated with smoothing of 10.3 cents and roll-off of 0.42) The greyscale indicates the spectral pitch distance with the reference triad (the darker the shade, the lower the distance and hence the greater the modelled affinity). A selection of major and minor triads are labelled—upper case for major, lower case for minor.

practice tonality) lie on the bottom-left to top-right diagonal. Chords above and below this are not major or minor, and in some of them the 'third' may be higher than both notes in the fifth (which is why I place 'third' in quotation marks to signal its unusual usage). For convenience, the central block represents the chord Cmaj. A selection of triads are labelled on the chart—upper case for major triads, lower case for minor. The chords are laid out in this way because this ensures their spatial distances correspond to their Euclidean voice-leading distances. For example, observe how the spatial distance between Cmaj and Emaj is twice the spatial distance between Cmaj and Dmaj.

The shade of each block indicates the spectral pitch distance (cosine) of its chord from the central Cmaj reference chord—darker is closer. For example, the shading of the Dmin block shows the spectral pitch distance between Cmaj and Dmin; the shading of the Fmaj block shows the spectral pitch distance between Cmaj and Fmaj. Spectral pitch distance is here calculated by embedding the first twelve partials (harmonics) of each triad tone as pitch classes into an absolute monad expectation tensor (a spectral pitch vector). This means each spectral pitch vector embeds a total of thirty-six pitch classes (i.e., $I = 36$). Pitch classes are used rather than pitches because I am interested in modelling

the perceived distances of chords regardless of whether any given voice moves up or down. For example, I want the spectral pitch distance between Cmaj and Fmaj, to reflect the distance between these two chords regardless of whether the root of the first chord ascends a fourth to the root of the latter or, instead, descends a fifth. The harmonics of each tone are weighted by $x_{\mathrm{w}}[i] = i^{-0.42}$ where i is the number of the harmonic, and the width (standard deviation) of the Gaussian smoothing is set to 10.3 cents. I use these parameter values because they are the optimal values for the model of the experimental data described in Chapter 4.

These spectral pitch distances can be used to model the perceived affinity between each of the triads and the reference Cmaj triad. Observe how there is a complex patchwork of differing distances across the voice-leading space. This model suggests that the triad pair {Cmaj, Dmin} has greater affinity than the neighbouring triad pair {Cmaj, Dmaj} (the rectangle labelled 'd' is darker than the rectangle labelled 'D'); the triad pair {Cmaj, Gmaj} has greater affinity than the neighbouring triad pair {Cmaj, G♭maj}; the triad pair {Cmaj, Emin} has greater affinity than the neighbouring triad pair {Cmaj, Emaj}; and so forth. Interestingly, it shows how chords with common tones and root movements by perfect fifth tend to have higher spectral similarity than chords without; that chords from the same diatonic scale tend to have higher spectral similarity than chords that are chromatically related. This corresponds to typical judgements of the affinity between chords, which suggests spectral pitch similarity can be used as an effective model of the affinities not just of successive tones, but also of successive chords. Indeed this hypothesis was experimentally tested in Milne (2009a): there was a correlation $r(24) = .78, p < .001$ between spectral pitch similarity and the ratings, by 32 participants, of the 'fit' of all 26 possible pairs of root-position major and minor triads.[17] A similar chart with a minor reference triad at the centre is given in Chapter 6.

17 The measure of spectral pitch similarity used in the cited thesis, termed *spectral response distance*, is subtly different to the spectral pitch similarity used here. As mentioned at the end Chapter 2, it was not a true metric.

I will revisit this chart in Chapter 6, where I show how it may be useful for explaining certain important aspects of tonal harmony.

3.6.2 *Temperaments*

Relative expectation tensors are useful to find effective *temperaments*. These are lower-dimensional tunings that provide good approximations of higher-dimensional tunings (Milne et al., 2008). The *dimension* of a tuning is the minimum number of unique intervals (expressed in a log-frequency measure like cents or semitones) that are required to generate, by linear combination, all of its intervals.

Many useful musical pitch collections are high-dimensional; for example, just intonation intervals and chords with frequency ratios 4:5:6 and 4:5:6:7 are three- and four-dimensional, respectively.[18] But lower-dimensional tunings (principally one and two-dimensional) also have a number of musically useful features; notably, they facilitate modulation between keys, they can generate scales with simply patterned structures (equal step scales in the case of 1-D tunings, well-formed scales in the case of 2-D tunings (Carey, 2007)), and the tuning of all tones in the scale can be meaningfully controlled, by a musician, with a single parameter (Milne et al., 2007).[19]

Given the structural advantages of low-dimensional generated scales, it is useful to find examples of such scales that also contain a high proportion of tone-tuples whose pitches approximate privileged higher-dimensional intervals and chords (such as just intonation). A familiar example is the 12-TET chromatic scale generated by the 100 cent semi-

18 *Just intonation* intervals and chords have low integer frequency ratios. They are typically thought to sound more consonant than the tunings used in 12-tone equal temperament.

19 *Well-formed scales* (Carey and Clampitt, 1989), or *MOS scales* (Wilson, 1975), are special cases of 2-D scales, each of whose generic intervals (as measured by number of scale steps) comes in no more than two specific sizes (as measured by a log-frequency unit like cents). In order to construct an MOS scale given a specific period and generator, the generator must be iterated precisely a number of times that yields a scale satisfying these requirements. The familiar (anhemitonic) pentatonic and diatonic scales are MOS scales with a period of 1200 cents and a generator of approximately 700 cents—the generator is iterated four times for the pentatonic scale, and two additional times for the diatonic scale. Numerous unfamiliar possibilities become available with non-standard period and generator tunings (Erlich, 2006).

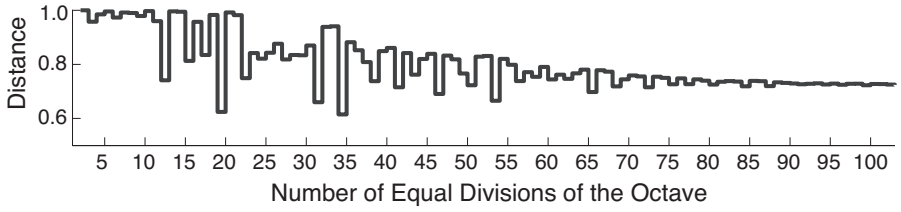


Figure 3.3: The cosine distance (on relative dyad expectation embeddings with a Gaussian smoothing kernel of 3 cents standard deviation) between a just intonation major triad $(0, 386.3, 702)$ and all n-tets from $n = 2$ to $n = 102$.

tone, which contains twelve major and twelve minor triads tuned reasonably close to their just intonation versions. Another familiar example is the meantone tuning of the diatonic scale, which is generated by a period (octave) of approximately 1200 cents and a generator (fifth) of approximately 697 cents—this scale contains three major and three minor triads whose tuning is very close to just intonation (closer than the 12-TET scale). There are, however, numerous alternative—and less familiar—possibilities.

Given a privileged pitch class collection embedded in an expectation tensor, it is easy to calculate its distance from a set of n-TETs (up to any given value of n).

Example 3.6.2. *1-D approximations to 4:5:6 (JI major triad).* The JI (just intonation) major triad contains all (and only) the common-practice harmonic consonances (i.e., the perfect fifth and fourth, and the major and minor thirds and sixths). It is, therefore, interesting to find tunings that produce simple scales containing lots of good approximations of these intervals. The just intonation major triad has pitches with frequency ratios of 4:5:6, hence it is three-dimensional (it is factorized by the three primes 2, 3, and 5). This means there is no equally tempered scale (which is one-dimensional, by definition) that can precisely match all of these just intonation intervals. However, certain equally tempered scale may provide reasonable approximations. Figure 3.3 shows the cosine distance between the relative dyad expectation tensor embeddings of the JI major triad and all n-TETs from $n = 2$ to 102.

Observe that the distances approach a flat line where increasing n is no longer beneficial, and that the most prominent minima fall at the familiar 12-TET and at other alternative n-TET's (e.g., 19-, 22-, 31-, 34-, 41-, 46-, and 53-TET) that are well-known in the microtonal literature.

A two-dimensional tuning has two generating intervals with sizes, in $\log(f)$, denoted α and β. All intervals in the tuning can be generated by α and β; that is, they have sizes $j\alpha + k\beta$ where $j, k \in \mathbb{Z}$. For example, all possible notes in the quarter-comma meantone tuning can be generated by an octave of 1200 cents and a generator of 696.58 cents (i.e., $\alpha = 1200$ and $\beta = 696.58$, so all intervals take the form $1200j + 696.58k$).

It is often convenient to create a scale from a β-*chain*, which comprises all pitches $j\alpha + k\beta$ where k is restricted to a limited range of successive integers. For example, a 19-tone β-chain might consist of the notes $j\alpha - 9\beta, j\alpha - 8\beta, \ldots, j\alpha + 8\beta, j\alpha + 9\beta$. Given an arbitrary set of higher-dimensional privileged intervals with a period of repetition ρ (typically 1200 cents), how can a β-chain with similarly sized intervals be found? In general, it is convenient to fix the tuning of α to ρ/n, for $n \in \mathbb{N}$, because this ensures the resulting generated scale repeats at the period whatever the value of β. So, once α is chosen, the procedure is to generate β-chains of a given cardinality and to iterate the size of β over the desired range. At each iteration, the cosine distance to the set of privileged intervals is measured using the relative dyad expectation embeddings. This is illustrated in the following example.

Example 3.6.3. *2-D approximations to 4:5:6 (JI major triad)—comparing smoothing widths*. Given a 19-tone β-chain with $\alpha = 1200$, we may wish to find the tunings of β that produce a large number of good approximations to the intervals found in a just-intonation major triad. The just intonation major triad is three-dimensional (as discussed in the previous example), while the β-chain is, by definition, two-dimensional. This means the latter can only approximate the former, never precisely match it. But we can iterate through values of β to find β-chains that contain large numbers of intervals that are close to the just intonation intervals.

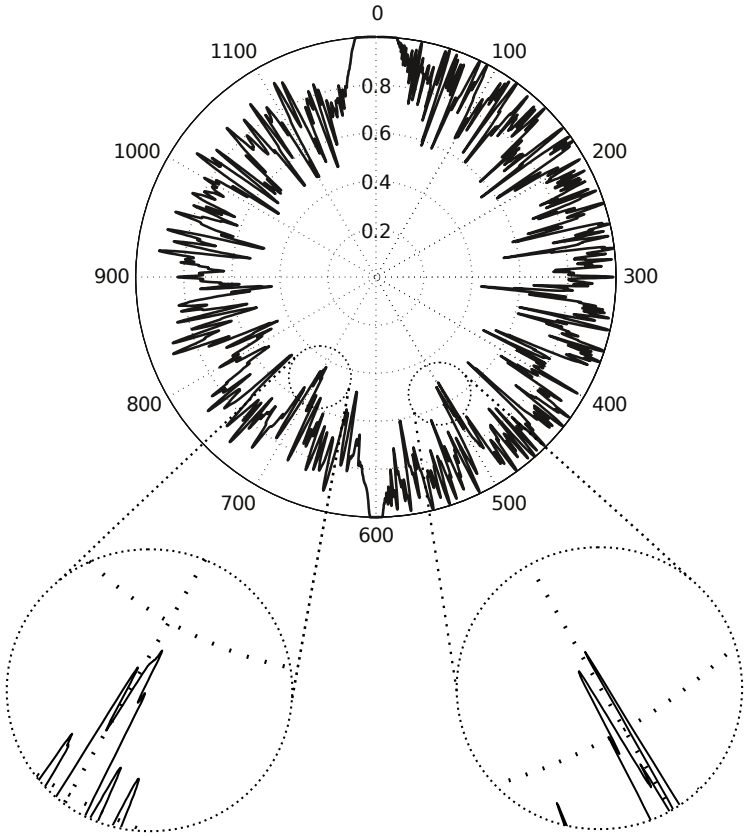


Figure 3.4: The cosine distance between relative dyad embeddings of a just intonation major triad $\{0, 386.3, 702\}$ and a 19-tone β-chain whose β-tuning ranges from 0 to 1,199.9 cents. The smoothing is Gaussian with standard deviations of 6 cents (left side), and 3 cents (right side). The two zooms show the distance minima occurring at the meantone (504 and 696 cents) and helmholtz (498 and 702) tunings, and how their relative levels change as a function of smoothing width.

Figure 3.4 shows the distance between the relative dyad embeddings of a just intonation major triad and 19-tone β-tunings ranging over $0 \leq \beta \leq 1199.9$ cents in increments of 0.1 cents ($\alpha = 1200$ cents). When using a single smoothing width, this type of chart is perfectly symmetrical about the centre line passing through 0 and 600 cents. This is because a β-chain generated by $\beta = B$ cents is identical to that generated by $\beta = \alpha - B$ (Milne et al., 2008). This means we can utilize such a chart to compare two different smoothing widths. In Figure 3.4, the right-hand side shows the effect of using a Gaussian kernel with a standard deviation of 3 cents; the left-hand side has a standard deviation of 6 cents.

Observe the following distance minima at different β-tunings: 503.8 cents corresponds to the familiar meantone temperament; 498.3 cents to the *helmholtz* temperament; 442.9 cents to the *sensipent* temperament; 387.8 cents to the *würschmidt* temperament; 379.9 cents to the *magic* temperament; 317.1 to the *hanson* temperament; 271.6 cents to the *orson* temperament; 176.3 cents to the *tetracot* temperament (the names for each of these temperaments has been taken from Erlich (2006)). It is interesting to note that the classic meantone tunings of approximately 504 (or 696) cents are deemed closer than the helmholtz tunings of approximately 498 (or 702) cents when the smoothing has 6 cents, and vice versa when the smoothing has a 3 cent standard deviation. In future experiments, the smoothing width could be used as a free parameter that is adjusted to best fit participants' ratings of how well various temperaments approximate just intonation.

Example 3.6.4. *2-D approximations to 4:5:6 (JI major triad)—comparing dyad and triad embeddings.* Figure 3.5 compares the distance between between a just intonation major triad and seven-tone β-chains (e.g., the notes $j\alpha - 3\beta, j\alpha - 2\beta, \ldots, j\alpha + 2\beta, j\alpha + 3\beta$, with β iterated over the sizes 0 to 1199.9 cents in increments of 0.1 cents) when embedded in relative dyad and relative triad expectation tensors. The left side shows triad embeddings, the right side shows dyad embeddings.

Observe that, for low cardinality generated scales (like this seven-tone scale) and a smoothing width of 3 cents, only a few tunings provide tone triples that are close to the just intonation major triad: the meantone generated scale ($\beta \approx 696$ cents) contains three major triads, the magic scale ($\beta \approx 820$ cents) contains two major triads, the porcupine scale ($\beta \approx 1,037$ cents) contains two major triads (but with less accurate tuning than the magic), the hanson scale ($\beta \approx 883$ cents) scale contains only one major triad (tuned extremely close to just intonation). As the cardinality of the β-chain is increased, the distances between the triadic embeddings approach those of the dyadic.

Example 3.6.5. *2-D approximations to 3:5:7 (7-limit Bohlen-Pierce triad).* The above two examples have used familiar tonal structures (the oc-

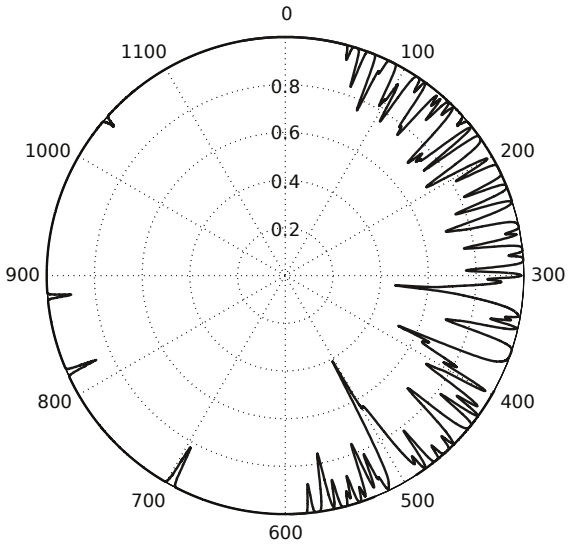


Figure 3.5: The cosine distance between relative dyad embeddings (right) and relative triad embeddings (left) of a just intonation major triad $\{0, 386.3, 702\}$ and a 7-tone β-chain whose β-tuning ranges from 0 to 1,199.9 cents. The smoothing is Gaussian with a standard deviation of 3 cents.

tave of 1200 cents and the major triad), but the methods are equally applicable to any alternative structure. One such is the Bohlen-Pierce scale, which is intended for spectra containing only odd numbered harmonics. It has a period of 3/1 (the 'tritave'), which is approximated by 1902 cents. The 3:5:7 triad, which is approximated by $\{0, 884.4, 1466.9\}$ cents, is treated as a consonance. Figure 3.6 shows the distance of a β-chain of 19 notes with $0 \leq \beta \leq 1901.9$ cents with a Gaussian smoothing of 3 cents standard deviation. The closest tuning is found at 439.5 cents, which is almost equivalent to $3 \times 1902/13$ and so corresponds to the 13-equal divisions of the tritave tuning suggested by Bohlen and Pierce (Bohlen, 1978; Mathews et al., 1984).

3.6.3 *Musical Set Theory*

In musical set theory, there is a rich heritage of measures used to model the perceived distances between pitch collections (e.g., Forte (1973); Castrén (1994); Buchler (2001); Kuusi (2001)). Expectation tensors can generalize traditional embeddings in a number of ways: (a) they model

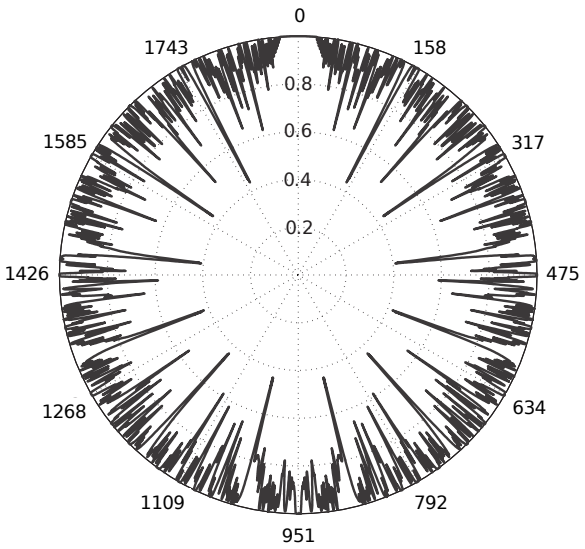


Figure 3.6: The cosine distance (using a Gaussian smoothing kernel with a 3 cents standard deviation) between a just intonation Bohlen-Pierce 'major' triad $\{0, 884.4, 1466.9\}$, with a period of 1902 cents, and a 19-tone β-chain whose β-tuning ranges from 0 to 1901.9 cents.

the inaccuracies of pitch perception; (b) they can embed pitch collections in any tuning (up to the pitch granularity determined by J); (c) they can meaningfully deal with pitch collections that contain duplicated pitches (such as when two voices play the same pitch); (d) they can be populated with pitches or pitch classes; (e) they can embed absolute or relative pitches or pitch classes; (f) they can generalize subset-class vectors, but utilize a principled indexing that does not rely upon Forte numbers.[20]

The relative dyad embedding is of the $\mathrm{T}_n\mathrm{I}$ type—that is, it is invariant with respect to transposition and inversion of the pitch collection it is derived from.[21] It is also invariant over Z-relations (Z-related collec-

20 A *set class* is an equivalence class for pitch class sets that differ only by transposition. *Forte numbers* are the numerical labels used by Forte to index set classes of all possible cardinalities (under the assumption of 12-TET, this ranges from the empty set up to the set of cardinality 12, and there are a total of 352 different set classes) (Forte, 1973). Given a pitch class set, its *subset-class vector* of cardinality n (also termed an *n-class vector*) indicates the number of occurrences of each set class of cardinality n, indexed by their Forte number (Kuusi, 2001).

21 In musical set theory, pitch class sets that are invariant with respect to transposition belong to the same T_n class. For example, $\{0, 4, 7\}$ and $\{1, 5, 8\}$ are in the same T_n class because the latter can be transposed down one semitone to make it equal to the former ($\{1-1, 5-1, 8-1\} = \{0, 4, 7\}$). Pitch class sets that are invariant with respect to both transposition and inversion belong to the same $\mathrm{T}_n\mathrm{I}$ class. For example, $\{0, 4, 7\}$ and $\{1, 4, 8\}$ are in the same $\mathrm{T}_n\mathrm{I}$ class because the latter can be converted into

tions, such as $\{0, 1, 4, 6\}$ and $\{0, 1, 3, 7\}$, have the same interval content but are not related by transposition or inversion (Forte, 1973)). Relative triad (and higher-ad) embeddings are invariant only with respect to transposition—that is they are of the T_n type (e.g., the inversion of a major triad is a minor triad and, although these two chords contain the same intervals, they have different embeddings in a relative triad matrix). When used with pitch class vectors, absolute embeddings have only period (octave) invariance; when used with pitch vectors, they have no invariances.

Example 3.6.6. *Distances between pc-sets related by Z-relation, inversion, and transposition*. Table 3.2 shows the cosine distances between the absolute and relative dyad and triad embeddings of pitch class vector $(0, 1, 4, 6)$, its Z-relation $(0, 1, 3, 7)$, its inversion $(0, 2, 5, 6)$, and its transposition $(1, 2, 5, 7)$.

It is reasonable to think that perceptions of pc-set similarity may be determined by both their absolute and relative pitch structures. To model this, pc-set similarity can be modelled as a linear combination of the distances between absolute and relative embeddings of differing orders. For example, adding relative dyad and absolute monad distances, gives a non-zero distance between pc-sets with differing interval content (like $(0, 1, 4, 5)$ and $(0, 1, 4, 6)$), but also takes into account their absolute pitches, thus ensuring $(0, 1, 4, 5)$ is closer to its transposition $(4, 5, 8, 9)$ than it is to its transposition $(2, 3, 6, 7)$ (e.g., adding the two distance functions, with no weighting, gives summed cosine distances of 0.533, 0.5 and 1, respectively).

3.7 DISCUSSION

In this chapter, I have presented a novel family of embeddings and metrics for modelling the perceived distance between pitch collections. The embeddings can be realized in a manner that conforms with established psychoacoustic data on pitch perception (through the use of

the former by inversion (which gives $\{-1, -4, -8\}$) and then transposition up eight semitones ($\{-1 + 8, -4 + 8, -8 + 8\} = \{0, 4, 7\}$).

Table 3.2: Cosine distances between a selection of pc-sets related by Z-relation, inversion, and transposition. The distances are calculated with four different types embedding.

		Z-relation	inversion	transposition
	$(0,1,4,6)$	$(0,1,3,7)$	$(0,2,5,6)$	$(1,2,5,7)$
Absolute dyad embeddings				
$(0,1,4,6)$	0	.833	.833	1
$(0,1,3,7)$	.833	0	1	.833
$(0,2,5,6)$	.833	1	0	.833
$(1,2,5,7)$	1	.833	.833	0
Relative dyad embeddings				
$(0,1,4,6)$	0	0	0	0
$(0,1,3,7)$	0	0	0	0
$(0,2,5,6)$	0	0	0	0
$(1,2,5,7)$	0	0	0	0
Absolute triad embeddings				
$(0,1,4,6)$	0	1	1	1
$(0,1,3,7)$	1	0	1	1
$(0,2,5,6)$	1	1	0	1
$(1,2,5,7)$	1	1	1	0
Relative triad embeddings				
$(0,1,4,6)$	0	1	1	0
$(0,1,3,7)$	1	0	.5	1
$(0,2,5,6)$	1	.5	0	1
$(1,2,5,7)$	0	1	1	0

Gaussian smoothing) and may be useful as components in broader models of the perception and cognition of music. Indeed, to model any specific aspect of musical perception, a variety of appropriate embeddings may be linearly combined, with their weightings, the weightings of the tone saliences (if appropriate), and the type of metric, as free parameters to be determined from empirical data.

The models demonstrated in this chapter differ from those of, for example, Krumhansl, Lerdahl, or those taking a neo-Riemannian or *Tonnetz*-based approach, because they are built from explicit psychoacoustic first principles (using Gaussian smoothing to model the frequency difference limen). Furthermore, unlike the traditional pitch embeddings used in set class theory, they are able to deal in a meaningful

way with non-standard tunings and when more than one tone plays the same, or a very similar, pitch.

This chapter has focused on expectation tensors, but the underlying pitch (class) response matrices can also be used to generate salience (rather than expectation) tensors: these give the probability of perceiving any given R-ad of pitches (rather than the expected number of tone-tuples perceived at a given R-ad of pitches). Developing computational simplifications for higher-order salience tensors is work that remains to be done.

The embeddings and metrics described in this chapter are also applicable to other domains: a tone, as defined at the start of this chapter, can be thought of as a member of a class of *discrete and linear stimuli*. A stimulus is *discrete* when it can be combined with other such stimuli, yet still be individually perceived (e.g., many tones may be sounded together, but still be individually heard; even the separate spectral pitches of a harmonic complex tone may be consciously perceived); a stimulus is *linear* when it can be characterized by a scalar that is the variable in a linear psychophysical function (e.g., a tone can be characterized by its log-frequency, which is linearly related to its perceived pitch height). In this generalized context, a *period* indicates the size—in the units of the psychophysical function—at which perceptual equivalence may occur (e.g., pitches that are octaves apart). These generalized definitions indicate how the same methods may be applied to the perception of any other (even non-auditory) discrete stimuli that can be transformed, with a link function, to make the psychophysical function linear. An obvious example is the perception of timing in rhythms: the 'physical' time of a percussive event is linearly related to the perceived time of the event, and a bar (or some multiple, or division, thereof) can be thought of as representing the period. In this context, the smoothing represents perceptual or cognitive inaccuracies in timing; for example, it might be possible to embed a rhythmic motif containing four events in a relative tetrad expectation matrix (in the time domain), and compare this with a selection of other similarly embedded rhythm patterns

to find one with the closest match (i.e., one that contains the greatest number of patterns that are similar to the complete motif).

In the subsequent two chapters, however, I focus my attention on the embedding of spectral pitches into absolute monad expectation tensors—spectral pitch vectors. In the next chapter, I use the cosine similarities of such vectors to model affinity of microtonally pitched tones with non-harmonic spectra. In the chapter after that, I use these cosine similarities of spectral pitch vectors to model Krumhansl's probe tone data, and to predict the tonal functions of pitches and chords in a variety of scales.

4

A MODEL OF MELODIC AFFINITY

In the previous chapter, I introduced a psychoacoustically derived method for embedding spectral (or virtual) pitches into a spectral (or virtual) pitch vector (virtual pitches were defined in Sec. 2.4). I suggested the cosine similarity of any two such vectors could be used to model the perceived affinity (the melodic analogue of consonance) of the tones or chords they embed. I also provided specific examples in Sec. 3.6.1. The principal aims of the experiment and models described in this chapter are: (a) to test the spectral pitch similarity model against experimental data; (b) to test whether it is modelling an underlying psychoacoustic process, rather than a learned response; (c) to determine the strength of the psychoacoustic effect modelled by spectral (or virtual) pitch similarity.

In addition to these aims, I also utilize cross-correlation between the spectral pitch vector of a tone and the spectral pitch vector of a *harmonic template* (defined in Sec. 4.1.1) to create a *harmonicity* model of the former's *toneness* (defined in Sec. 4.1.2). This is also tested as a model for affinity. Furthermore, I also embed virtual pitches, instead of spectral, to see if virtual pitch similarity can also provide an effective model of affinity.

As I demonstrate later, the experimental data collected for this chapter confirm that spectral pitch similarity and harmonicity are effective models of affinity, and combining them produces a model with a large effect size that is highly significant.[1] Virtual pitch similarity performs slightly worse than spectral pitch similarity, but not significantly so. The data also support the hypothesis that spectral pitch similarity is a

1 *Effect size* is any measure (e.g., correlation) of the degree of association between two variables (e.g., a model's predictions and the data it is modelling). Its value is important because, given enough data, an effect size that is vanishingly small (i.e., unimportant) can still be statistically significant.

model of an innate psychoacoustic process (it is not learned from exposure to a musical corpus), and that the size of this psychoacoustic effect is medium and also highly significant.

It is widely accepted that the spectra of simultaneously played tones have an impact on their perceived consonance and dissonance (the *spectrum* of a tone is the set of amplitudes and frequencies of all its partials). For example, the Helmholtz (1877) tradition ascribes *dissonance* to partials in the combined sound that are close enough in frequency to interfere with each other, thereby creating rough-sounding rapid beating. Meanwhile, the Stumpf (1890)[2] tradition ascribes *consonance* to *fusion*: when partials have frequencies close to a harmonic series, they are perceptually fused into a smaller number of virtual pitches, thereby simplifying the percept (virtual pitches are inferred by the auditory system and correspond to the frequencies of candidate fundamentals). As suggested by many researchers, it is plausible that both processes—in addition to familiarity—contribute to the final consonance/dissonance of a chord (e.g., Bregman (1990); Huron (1991); Lerdahl (2001); McDermott et al. (2010); Parncutt and Hair (2011)).

However, the suggestion that the spectral content of *melodically played* (i.e., successive) tones may have an impact on their *affinity* (their perceived degree of 'fit', 'in-tuneness', and 'similarity') is less well known and typically ignored in favour of learning models (e.g., tones separated by more familiar intervals fit better than those separated by less familiar intervals) such as those developed by Krumhansl (1990) and Pearce and Wiggins (2006). However, Terhardt (1984) suggested there are at least two mechanisms by which spectral content can affect the affinity of two tones: the commonalities of their spectral pitches, and the commonalities of their virtual pitches—any given notated pitch may produce a multiplicity of both spectral and virtual pitches. In this dissertation, I use the terms *spectral pitch similarity* and *virtual pitch similarity*; this is because, in my models, I treat each spectral and virtual pitch as a random variable and model the similarity of collections of such pitches with

2 As cited in (Green and Butler, 2002).

a similarity measure such as the cosine. Spectral pitch similarity—the cosine similarity of two spectral pitch vectors—was introduced in Section 2.5. Virtual pitch similarity is the cosine similarity of two virtual pitch vectors. *Virtual pitch vectors* are identical to spectral pitch vectors except they embed all and only the virtual pitches that have been generated by an additional model of virtual pitches (the spectral and virtual pitch models used in this chapter are fully described in Sec. 4.1.1 and App. D). I use the term *pitch similarity* when I wish to be non-specific as to whether the mechanism is spectral pitch similarity, virtual pitch similarity, or both.

Both mechanisms seem plausible because harmonic complex tones an octave, perfect fifth, or perfect fourth apart have high spectral and virtual pitch similarities, and these intervals are typically considered to have a high melodic affinity (as historically evidenced by Tenney (1988)). Furthermore, these processes were experimentally tested by Parncutt (1989) whose empirical data showed that melodic unisons, octaves, and perfect fifths and fourths were rated as significantly more 'similar' than some of their neighbours a semitone away—the minor second, major seventh, and tritone.

There is also a teleological argument that naturally associates pitch similarity with affinity. Successive tones with similar pitch contents clearly exhibit greater simplicity and continuity, both of which are important components of *Prägnanz* (good form). In Gestalt perceptual theory, the *law of Prägnanz* asserts that perceptual elements are grouped due to them forming part of a larger well-formed pattern. By playing successive tones that share some of their spectral or virtual pitch content, a composer or performer is suggesting a patterning and orderliness that may help bind the melodic notes into a greater whole (such as a melodic stream where notes are heard to be, in some sense, connected rather than disjointed). Over the continuum of all possible intervals sizes, most intervals have low pitch similarity (e.g., see Fig. 4.11); those with high similarity are rarities. Privileging them, therefore, represents a clear aesthetic decision.

The above suggest a causal relationship from the pitch similarity of two tones to their perceived affinity, but there is a complication for any experiment designed to test such a relationship: most listeners are familiar with Western music, which privileges—by prevalence and metrical weight—certain melodic intervals over others (e.g., seconds are preferred to sevenths; perfect fifths and fourths to tritones; and so forth), and which privileges certain spectral tunings (the harmonic complex tones produced by most Western instruments and the human voice).[3] This means that, in an experimental setting, participants' responses to different intervals are likely to be a function of their familiarity in addition to any innate psychoacoustical or cognitive processes (e.g., in Western music, the perfect fifth is more prevalent than the tritone, and this contributes to its greater affinity). This would appear to make it difficult to determine whether perceived melodic affinity is due to bottom-up (e.g., psychoacoustic) processes, due to learning, or to some combination of the two (and what the relative strengths of their contributions are).

To overcome this problem, I conducted an experiment—of the type suggested at the end of Section 2.3.4—which uses randomly generated microtonal melodies, each containing 16 eighth-notes (two bars of 4/4). The timbres used were synthetic and contained non-harmonic spectra. This means that both the melodic intervals and the timbres were unfamiliar. Each microtonal melody was played with two different timbres, and participants were asked to choose in which timbre the tones fitted together the best. In all cases, one timbre had partials tuned to match the melody's underlying tuning, the other did not (in a *matched timbre*, the tunings of the partials are adjusted so they have frequencies that match those used in the underlying scale—this is explained, in more detail, in Section 4.2.2). When a timbre is matched, the resulting melody tones typically have greater pitch similarity—both spectral and virtual—than when the timbre is unmatched. Asking participants to choose between two different timbres, each playing precisely the same melody, effectively removes the confounding influence of inter-

3 A *spectral tuning* is the set of frequencies of all partials relative to the lowest.

vallic familiarity. Removing this important top-down effect enables the specific influence of spectral pitch similarity to be more clearly observed and modelled.

It is, however, not possible to remove all possible top-down influences. For example, in Western music, high affinity intervals are typically high in pitch similarity (notably, the unison, octave, perfect fifth and fourth). It may be that we carry over this learned association to unfamiliar microtonal tunings and timbres. Clearly this top-down process can work in tandem with the bottom-up process I have proposed—indeed, it is a nice example of a positive feedback loop like those discussed in Section 2.3.3. However, in this experiment, I do remove the most obvious top-down component—that affinity is a direct function of familiarity.

The experiment allows two principal hypotheses to be tested. The first hypothesis is that melodic affinity increases when its tones have greater pitch similarity. The second hypothesis is that melodic affinity increases when the tones, themselves, have partials that are approximately harmonically related (they have higher *harmonicity*). Due to the experimental design, we can be sure that any observed association between pitch similarity and affinity is not due to familiarities with different interval sizes (a top-down process). For any observed association between harmonicity and affinity, the experimental design does not eliminate the possibility that familiarity with harmonic complex tones (a top-down process) is a cause.[4]

In subsequent sections, I show that the results of the experiment bear out both hypotheses: participants chose the (typically higher pitch similarity) matched timbre significantly more often than the (typically lower pitch similarity) unmatched timbre (61% of occasions, $p < .001$ exact binomial test) (Section 4.3.1). Furthermore, values generated by nonlinear models of the spectral pitch similarities and harmonicities for every stimulus were used as predictors in a logistic regression on the data (the probabilities of choosing the matched timbre). The re-

4 Familiarity with harmonic complex tones is effectively universal because the human voice has this spectral tuning.

sulting fit was good, and cross-validation demonstrates the model (and each predictor) generalizes well and is statistically significant. To provide easy-to-understand statistics, the same spectral pitch similarity and harmonicity values were also used as predictors in a linear model of the log-odds of choosing the matched timbre, and this resulted in $R^2 = .59$, with standardized coefficients of $.38$ for spectral pitch similarity, and $.70$ for harmonicity (Section 4.3.2). Of the pitch similarity models, spectral pitch similarity performed slightly (nonsignificantly) better than virtual pitch similarity, which is the reason I focus on the results of the former model.

This has some important implications (which are explored in Section 4.4). Firstly, it demonstrates that spectral pitch similarity can be used as a tool to widen the scope of psychoacoustic modelling from simultaneously played tones to melodically played (successive) tones. In particular, it should be possible to model the affinity of tones and chords in familiar Western scales by their spectral pitch similarity—such models may provide explanations for why certain scales are more common than others. Secondly, although it is well established that matching partials to an underlying scale-tuning can minimize sensory dissonance (Sethares, 2005), the results show this procedure has the potential to make microtonal melodies sound more in-tune (have greater affinity), even in the absence of chords.

In Section 4.1, I explain my model and how the experiment is designed to minimize the causal impacts of nurture processes. I describe the experimental method in Section 4.2, provide the results in Section 4.3 and, in Section 4.4, I examine some of the implications of the research.

4.1 THE MODEL OF MELODIC AFFINITY

In the experiment, which is fully described in Section 4.2, participants were asked to choose which of two timbres made the tones in a microtonal melody have the greatest 'affinity' (which was explained to them

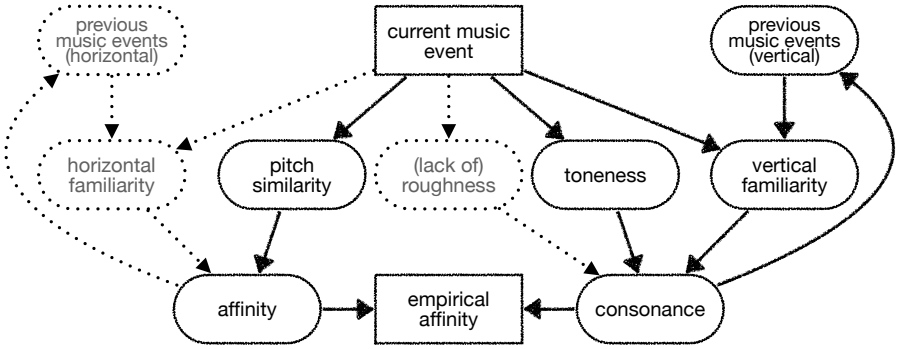


Figure 4.1: The full model of empirical affinity (i.e., affinity as reported by participants). Physical stimuli are in the top row, mental processes in the middle row, and subjective mental states, and empirical reports thereof, in the bottom row. Mental processes with an arrow from 'previous music events' are nurture processes (i.e., horizontal and vertical familiarities); those without, are nature processes (i.e., pitch similarity, roughness, and toneness). In the experiment described later, the impact of horizontal familiarity on affinity, and roughness on consonance, is minimized; for this reason, these processes (and submodels thereof) are not included in the final model of the experimental data, which is why they are greyed out and their causal paths are dashed.

as the degree to which successive tones 'fit', sound 'in tune', or 'unsurprising'). This process was repeated for a wide variety of differing melodies and timbres.

My full model of melodic affinity is shown in Figure 4.1. I hypothesize that participants' reports of affinity are due to two main causes: (a) the affinities of successive tones—labelled 'affinity' on the bottom row of Fig. 4.1; (b) the *inherent consonance* of each tone (i.e., the extent to which each tone's partials are harmonic, do not beat, and are familiar)—labeled 'consonance' on the bottom row of Fig. 4.1). That is, even though participants are asked to report solely on the affinity of successive tones, we expect them, mistakenly, also to respond to the inherent consonance of each individual tone (this is an example of an *untargeted* mental state, as discussed in Section 2.3.2). This is plausible because if a participant hears every individual tone as dissonant, the overall stimulus is somewhat unpleasant and is, therefore, unlikely to get a high rating for affinity (this was anticipated before the experiment was conducted).

As outlined in Section 2.2, I also hypothesize that affinity and consonance are monotonic functions of both bottom-up processes (pitch similarity for affinity, and roughness and toneness for consonance) and top-down processes (horizontal familiarity for affinity, and vertical familiarity for consonance). These are shown in the middle row of Figure 4.1. As defined earlier, *pitch similarity* is the similarity of the spectral or virtual pitches of any two successively played tones or chords; *roughness* is the perceptual correlate of the beating produced by partials close in frequency; *toneness* refers to the perceived extent to which a tone produces a single unambiguous pitch; *horizontal familiarity* refers to a listener's familiarity with non-simultaneous pitch intervals (or sequences of such intervals); *vertical familiarity* refers to a listener's familiarity with simultaneously played pitch intervals and chords (*horizontal* and *vertical* are used in the sense of musical notation—horizontal events occur across time, while vertical events occur simultaneously). In Sections 4.1.1 and 4.1.2, I give full definitions of these processes and also provide simple *submodels* of each of them. In Section 4.3.2, I show how these submodels are combined into the full model of the experimental data.

As described in Section 4.1.3, I have designed the experiment to minimize the impact of horizontal familiarity on affinity, and roughness on consonance, so these processes (and submodels thereof) are not included in the final model, which is why they are greyed out in Figure 4.1. The purpose of minimizing horizontal familiarity is to allow the bottom-up effect of pitch similarity to be more cleanly assessed (as described in Section 2.3.4, an experimental intervention like this allows the effect sizes of nature processes to be more fully disambiguated from those of nurture processes). The purpose of minimizing the impact of roughness is to simplify the final model required. Furthermore, for the stimuli used in the experiment, my submodels of affinity and consonance are not correlated. This ensures I can distinguish the causal strength of pitch similarity from the combined causal strengths of toneness and vertical familiarity. I cannot, however, distinguish the relative

strengths of toneness and vertical familiarity; this is because they are both modelled by harmonicity, as discussed in Section 4.1.2.

4.1.1 *Affinity: Pitch Similarity and Horizontal Familiarity*

I hypothesize that two harmonic complex tones will have: (a) higher affinity if more of their higher-amplitude partials align (e.g., the third partial of C4 has almost the same log-frequency as the second partial of G4; the sixth partial of C4 has almost the same log-frequency as the fourth partial of G4; and so on); (b) lower affinity if fewer of their higher-amplitude partials align (e.g., C4 and D♭4, which have few high-amplitude partials in alignment). Harmonic complex tones whose fundamentals have frequencies that are close to low-integer ratios (e.g., 3/2) have a greater number of coinciding partials: more precisely, for an interval with frequency ratio p/q, the ratio of the number of coinciding partials to all partials is given by $2/(p+q)$ (Milne, 2009a; Juhász, 2012). This gives a useful rule-of-thumb, but it cannot account for minor deviations from precise integer ratios (e.g., the 12-tone equal temperament fifth is 2 cents narrower than 3/2 and, in a typical performance, pitches frequently deviate from any predefined norm), nor can it take account of the amplitudes of the partials. For this reason, a more precise psychoacoustically-derived model is desirable. This can be provided by *spectral pitch similarity*, which is the cosine similarity (or other similarity measure) between the *spectral pitch vectors* of the partials found in each tone.

As described in Chapter 3, a *spectral pitch vector* comprises a high number (typically, thousands) of elements. The index value of each element indicates a specific pitch, while the value of each element indicates the expected number of partials heard at that pitch. The pitches are finely-grained and the index number of each element is proportional to its log-frequency—so the vector's first element might correspond to a MIDI note number of 60 (middle C), the second element to 60.01 (1

cent above middle C), the third element to 60.02 (2 cents above middle C), and so forth.

In the model used here, the values in the spectral pitch vector are a function of two parameters: *spectral roll-off* ρ, and *smoothing width* σ. These parameters are defined in detail in Section 4.3.2.1, but I will briefly describe them now. The former parameter models the relative importance of lower and higher partials in a tone. It does this by giving a weight of $1/i^{\rho}$ to each partial i, which means that as ρ increases the weights of the higher partials decrease. The latter parameter models inaccuracies of pitch perception. As described in Chapter 3 and Appendix A, this is achieved by convolving, in the log-frequency domain, the stimulus' frequency components with a discrete and truncated approximation of a normal distribution with a standard deviation of σ.[5] The effect of the smoothing is illustrated in Figure 4.2.

As described in Section 3.5, I model the perceived affinity of any two complex tones (or chords) by taking the cosine similarity of their spectral pitch vectors (any alternative similarity measure could be substituted). The above procedures are mathematically defined in Appendix D—specifically (D.1–D.3) and (D.6). As I will explain in Section 4.3.2.1, I used the spectral pitch similarities of the intervals and timbres in my experimental stimuli to model the responses given by participants.

I also hypothesize that another, related, bottom-up cause of affinity is *virtual pitch similarity*, and this can be modelled by the cosine similarity (or other similarity measure) of *virtual pitch vectors*. Virtual pitch vectors represent the virtual pitches, and their weights, that may be heard in response to a complex tone. The auditory system is thought to generate multiple virtual pitches that correspond to the frequencies of possible (candidate) fundamental frequencies (Terhardt et al., 1982; Parncutt, 1988). For instance, a harmonic complex tone (i.e., a tone whose partials are all integer multiples of a fundamental frequency) is typically

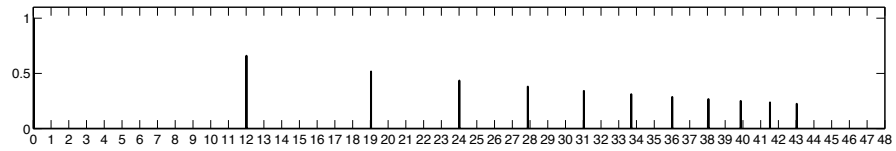
5 For random variables X and Y, the probability distribution of their sum $X + Y = Z$ is given by the convolution of their respective probability mass functions; that is, $p_Z(z) = p_X(x) * p_Y(y)$. This smoothing, therefore, represents the 'true' log-frequency stimulus plus normally distributed random deviations.

heard as having a principal pitch that corresponds to its fundamental frequency (this occurs even when the fundamental is removed from the stimulus), but virtual pitches may also be heard at subharmonics below this fundamental. A chord containing many harmonic complex tones may produce a variety of competing virtual pitches, with the root of the chord typically being the strongest (Parncutt, 1988). Similarly, in an inharmonic timbre like a bell, there may be several, very salient, competing virtual pitches.

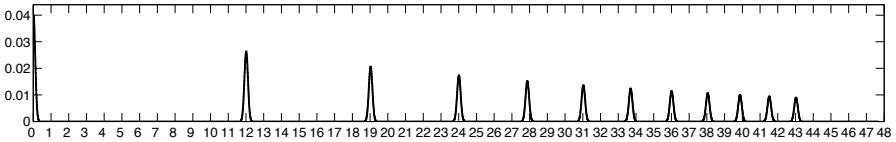
There are many possible ways to model the virtual pitches produced by a spectrum. I have chosen a simple approach (related to that introduced by Brown (1992)), which is to cross-correlate the spectral pitch vectors of the timbre and a *harmonic template*.[6] The latter can be thought of as a template harmonic complex tone that resides in long-term memory, and against which any steady-state timbre can be compared in order to find candidate fundamentals. For example, if the tone has a perfectly harmonic spectrum, this model will predict strong virtual pitches at the fundamental and harmonics above and subharmonics below the fundamental, and also at octaves of these (i.e., perfect fifths and fourths above and below). Figure 4.3 illustrates the virtual pitches generated by a harmonic complex tone using this model. As with the spectral pitches, I model the perceived affinity of any two complex tones (or chords) by taking the cosine similarity of their virtual pitch vectors. The processes described in this paragraph are mathematically defined in Appendix D—specifically (D.4) and (D.8). As I will explain in Section 4.3.2.2, I used the virtual pitch similarities of the intervals and timbres in my experiment to predict participants' responses.

For harmonic complex tones, the intervals between their virtual pitches are the same as those found between their spectral pitches (they just have different weights); this means that, for harmonic complex tones, spectral pitch similarities and virtual pitch similarities over dif-

6 Cross-correlation is a sliding dot product between two vectors—the first element of the cross-correlation vector is the dot product of the two input vectors, the second element of the cross-correlation vector is the dot product where the second vector has been shifted one step to the left, the third element of the cross-correlation is the dot product where the second vector is shifted two steps to the left, and so on.



(a) Unsmoothed spectrum.



(b) Smoothed spectrum.

Figure 4.2: The effect of smoothing (convolving) a spectrum with a discrete approximation of a normal distribution with a standard deviation of 10 cents.

ferent intervals can be almost perfectly correlated. I had expected that the use of non-harmonic spectra would substantially reduce this correlation, but this wasn't the case. When separately optimized to the data (as described in Sec. 4.3.2.4), the virtual pitch similarities and spectral pitch similarities have a correlation of $r(110) = .95$, which means they are effectively identical for modelling purposes; either could serve equally well as a model for pitch similarity, but both cannot be usefully included in the same model because estimates of their relative importance will be highly unreliable due to multicollinearity. For much of the following, I focus my attention on spectral pitch similarity because its model is simpler (it does not require cross-correlation with a harmonic template to generate pitches that are not in the stimulus); furthermore, a model using spectral pitch similarity was found to be have a marginally better fit to the data than a model using virtual pitch similarity (see Section 4.3.2.4).

Horizontal familiarity can be modelled by the probability of intervals in a corpus representing the musical system under consideration and the listener's experience. A model for horizontal familiarity is not required in this experiment (as explained in Section 4.1.3), so an explanation of the precise mechanics of calculating musical prominence is not given here, but a good example of this methodology can be found in Pearce and Wiggins (2006).

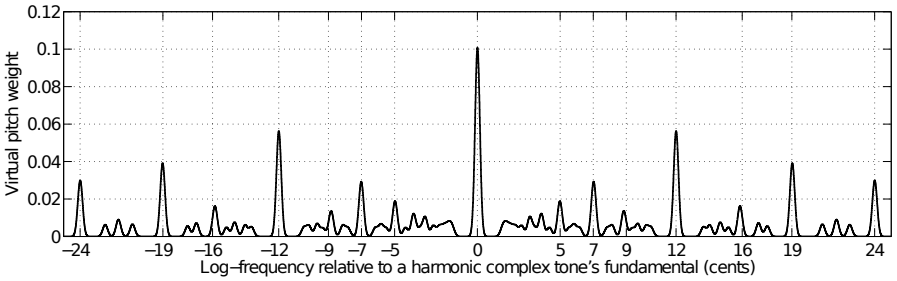


Figure 4.3: Virtual pitch weights for a harmonic complex tone as modelled by cross-correlating the spectral pitch vector in Figure 4.2b with a harmonic template.

4.1.2 *Consonance: Roughness, Toneness, and Vertical Familiarity*

Roughness is a perceptual/cognitive attribute that quantifies the perceived roughness or perceptual confusion of a tone (or chord) as a function of the interference (e.g., beating) caused by more than one partial occupying the same auditory filter.[7] It can be modelled by combining the sensory dissonances between all possible pairs of simultaneously occurring partials in a spectrum. The sensory dissonance of a single pair of partials can be calculated with a model of empirical data recording dissonance as a function of the frequency distance of the two partials and the size of the critical bandwidth at that frequency range—such as the Plomp and Levelt (1965) data as parameterized by Sethares (1993). These pairwise dissonances can be combined by a simple summation, a more complex power law, or a vector norm (see, e.g., the methods reviewed by Mashinter (2006)). This is a well-established type of model for perceived roughness or sensory dissonance (e.g., Plomp and Levelt (1965); Kameoka and Kuriyagawa (1969); Sethares (2005)).

Unsurprisingly, I hypothesize that consonance is a monotonically decreasing function of roughness (i.e., as roughness goes up, consonance goes down); I also hypothesize that this, in turn, will affect judgements of affinity. In my experiment, however, I control for roughness so I do not model this process in my final model.

7 The mammalian basilar membrane, within the cochlea, is typically modelled by large number of band-pass filters—each with a different centre frequency—termed *auditory filters*.

Toneness is a term defined by Huron (2001, p. 7) as the 'clarity of pitch perception' and the extent to which 'certain sets of pure tones may coalesce to form a single auditory image—as in the case of the perception of a complex tone'. This is, in part, a function of the frequency of the fundamental (e.g., tones between about 5 and 20 kHz are audible but do not produce a strong pitch sensation) but, for complex tones, it is also a function of the frequency relationships of their partials. As discussed above, harmonic complex tones have a relatively strong and unambiguous pitch, whereas tones with non-harmonic partials may have ambiguous or multiple pitches (like bells), or no clearly discernible pitch (like drums or cymbals). I hypothesize that toneness causally increases the perceived consonance of a timbre and, hence, the affinity of different tones using that timbre.

I model toneness by measuring the *harmonicity* of the timbre, which is the degree of similarity between the tone's spectrum and a harmonic template. We cannot know, a priori, what the pitch—if any—of an arbitrary inharmonic complex tone is, so harmonicity is modelled by cross-correlating the spectral pitch vector of the tone's spectrum and the spectral pitch vector of a harmonic template (as described in Section 4.1.1) and taking the maximum value of the resulting vector. This gives the value of the greatest possible (un-normalized) correlation of the tone's spectrum to the template, and the pitch at which this occurs. This process is mathematically described in Appendix D—specifically (D.5). As I will explain in Section 4.3.2.3, I used the harmonicity of the timbres in my experiment to model participants' responses.

Vertical familiarity is the extent to which an arrangement of simultaneous tones or partials is familiar. This can apply to the familiarity of different chords, and to the familiarity of different steady-state timbres. In the context of this experiment, only the latter is relevant because only melodies, not chords, were used; furthermore, the only aspect of the timbre that changed was the tuning of the partials. The familiarity of any given spectral tuning can be modelled by its prevalence in the musical system under consideration (see Section 2.3.2). However,

rather than carry out a detailed statistical analysis of Western music, I make the simple observation that most Western instruments, and the human voice, produce tones with harmonic spectra (or approximately harmonic spectra), and that such instruments have a privileged status in Western music. Hence, I model vertical familiarity with harmonicity.[8] I hypothesize, therefore, that harmonicity is monotonically related to vertical familiarity, hence to consonance, and hence to affinity.

This means that harmonicity is used to model both toneness and vertical familiarity so, in this experiment, I cannot distinguish between these two mental processes. To distinguish between them would ideally require participants who are familiar with non-harmonic timbres—for example, listeners from a gamelan tradition—because this would allow the models of toneness and vertical familiarity to de different (i.e., their predictions would be uncorrelated).

4.1.3 *Minimization and Control of Confounding Variables*

The principal purpose of the experiment described in this chapter is to examine whether participants' judgements of affinity are influenced by the nature process of pitch similarity; this is facilitated by minimizing the impact of the top-down process of horizontal familiarity. I am also interested in exploring the extent to which the harmonicity of the tones' spectra influences the affinity between them; this is facilitated if I minimize the impact of the bottom-up process of roughness. Furthermore, in order to ensure we can reliably distinguish between the impacts of pitch affinity (due to spectral and/or virtual pitch similarities) and consonance (due to harmonicity and vertical familiarity), I must ensure they are not strongly correlated; if they were, it would not be possible to reliably separate out their individual influences (due to their multicollinearity). I also seek to minimize the influence of any unforeseen confounding variables.

8 For participants familiar with a musical culture such as gamelan, where pitched instruments are typically inharmonic, a different model of vertical familiarity may be appropriate.

The experimental methods to achieve this are explained in the remainder of this section, but it is necessary first to provide a very brief introduction to the experimental procedure (the procedure is fully described in Section 4.2.3). Participants were played 60 different randomly generated melodies. Each melody was played in one of eleven different equal temperaments: 3-TET, 4-TET, 5-TET, 7-TET, 10-TET, 11-TET, 12-TET, 13-TET, 15-TET, 16-TET, and 17-TET. (An n-tone equal temperament—or n-TET—has n equally sized steps per octave; a familiar example being 12-TET.) For each melody, the participant could switch between two different timbres: one timbre was matched to the underlying tuning; the other was unmatched (its partials were matched to a different n-tet randomly chosen from the same list). The participant was asked to pick the timbre where the different notes of the melody had the greatest affinity.

4.1.3.1 *Minimizing the influence of horizontal familiarity and roughness*

Minimizing the influence of horizontal familiarity is achieved by asking participants to compare paired stimuli that have the same melody (and hence equivalent horizontal familiarities), but whose tones have differing timbres (and hence differing spectral pitch similarities).[9] In one stimulus, the tones' partials are tuned to match the underlying scale tuning so as to maximize spectral similarity; in the comparison stimulus, the tones' partials are not tuned to match the underlying scale tuning. This means that, for every pair of stimuli, one of the stimuli typically contains intervals with greater spectral pitch similarity, the other typically contains intervals with lower spectral pitch similarity; but both melodies are, in all other respects, identical. In other words, spectral pitch similarity varies between the two stimuli while horizontal familiarity remains fixed.

I seek to minimize the impact of roughness on participants' responses by making the roughnesses of both stimuli in each pair (matched and

9 It is possible that changes in spectral tuning may change the perceived pitch of the tones, but such changes should be uniform across differently pitched tones (so perceived interval size should be consistent).

unmatched timbres) as low as possible. This minimizes the differences between their roughness levels, and it also minimizes the possibility of the tones' roughnesses (inherent dissonances) dominating, or distracting from, the perception of the affinities between successive tones (which is the experimental target).

To achieve this, both matched and unmatched timbres are tuned to n-tone equal temperaments with low values of n (in this experiment, $n \leq 17$). Local minima of roughness (as calculated by Sethares' (2005) model) occur at such spectral tunings because no two partials can be closer than $1200/n$ apart and most of them will be at intervals of sizes $1200\,m/n$, where m is an integer greater than 1. This tends to ensure most partials are more distantly separated than the quarter of critical bandwidth that is maximally dissonant (Plomp and Levelt, 1965).[10] The relationship between spectral tuning and roughness is suggested in Figure 4.4. Here, a continuum of generated spectral tunings is shown (partials have been tuned to a linear combination of a period of 1200 cents and a generator with a tuning ranging between 680 and 725 cents). Note the deep and narrow minima at the labeled low-n n-TET tunings—these spectral tunings precisely correspond to those I used for the matched timbres at those tunings (shown in Table 4.2). The continuum of tunings shown here does not show all possible dimensions of spectral tunings (that would require an $(m+1)$-dimensional visualization, where m is the number of harmonics), but it does provide a glimpse of one such continuum.

I assume this technique is sufficiently effective at reducing the influence of roughness on judgements of melodic affinity that the variable of roughness (and its model) can be removed from my overall model of participants' responses.

4.1.3.2 *Ensuring harmonicity and pitch similarity are not correlated*

As described in Section 4.1.2, I model the toneness and vertical familiarity of a tone by the harmonicity of its spectrum. The selection of n-

10 For tones around D4 (294 Hz), the maximally dissonant log-frequency difference is approximately 80 cents, which is 1/15 of an octave.

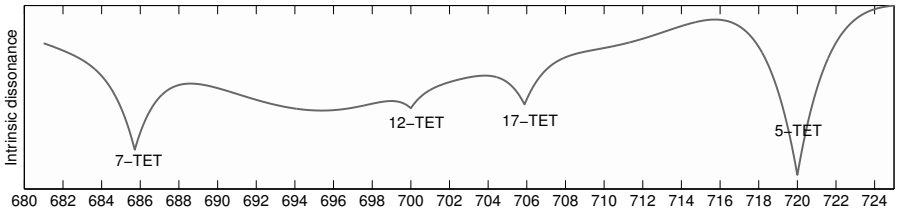


Figure 4.4: Inherent roughness—modelled with Sethares' (2005) routine—over a continuum of generated tunings (the generator is a fifth-like interval ranging from 680 to 725 cents). This tuning continuum includes a variety of n-TETs, only a few of which have a low value of n (these are labelled). Note that there is a broad local minimum of sensory dissonance when the partials are close to harmonic (approx. 696 cents), and narrow local minima at low-n n-TETs.

TETs used in the experiment produces a wide variety of timbres that are mostly rather inharmonic in quality (their harmonicity is low), though 12-, 15-, and 17-TET have somewhat greater harmonicity. Importantly, under the models described above, the differences between the harmonicities of the matched and unmatched timbres are not correlated with the differences between their spectral pitch similarities, so it is meaningful to regress measured affinity on both spectral pitch similarity and harmonicity so as to gauge their relative importance. For example, when the melody is in 12-TET and the unmatched timbre is in 5-TET, the matched timbre has greater spectral similarity and greater harmonicity; conversely, when the melody is in 5-TET and the unmatched timbre is in 12-TET, the matched timbre has greater spectral similarity but now it has lower harmonicity. The resulting model, therefore, takes the form shown in Figure 4.5.

4.1.3.3 *Controlling for unknown factors*

There may be a variety of unanticipated factors that affect judgements of affinity. In order to minimize their influence, the experiment was randomized in many different respects (as described in more detail in Section 4.2). Eleven different tunings were used and, for each pair of stimuli that participants were asked to rank, the scale tuning and the spectral tuning (of the unmatched timbre) were independently and randomly selected for each participant. Furthermore, for each pair of stimuli, the tones played, their articulation, and the overall tempo, were

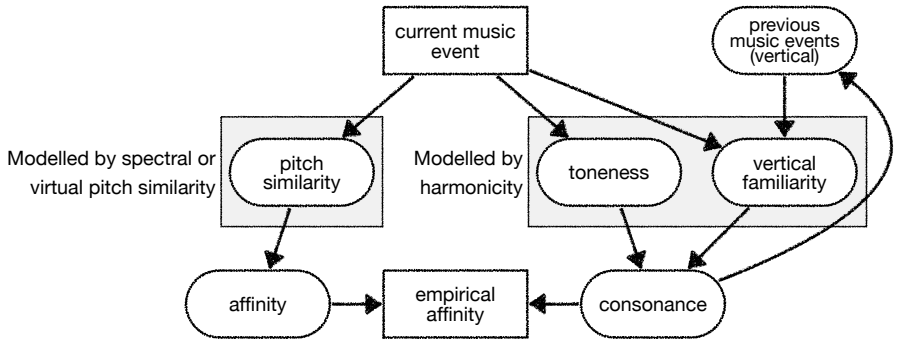


Figure 4.5: The final model of the empirical data—affinity is modelled by either spectral or virtual pitch similarity, consonance by harmonicity. The empirical data (participants' responses) is modelled as a combination of affinity and consonance. Pitch similarity and toneness are nature processes, vertical familiarity is a nurture process (by definition, because one of its inputs is 'previous music events'—see Sec. 2.3.2)

generated randomly (within a set of musically sensible constraints, see Section 4.2.2). Participants had a wide range of musical abilities and experience.

4.2 METHOD

4.2.1 *Participants*

Forty-four academic and non-academic university staff and graduate students participated in the experiment (25 male, 19 female, mean age 37.4 years, standard deviation 11.1 years), and no reimbursement was given. Eleven claimed to have had no musical training or ability; 12 to have had basic musical training or ability (Associated Board of the Royal Schools of Music Grades 1–4, or similar qualification or experience); 14 to have had intermediate training or ability (Grades 5–7, or similar); 7 to have had advanced training (Grade 8 or higher, or similar). The average level is, therefore, somewhere between basic and intermediate, and the overall distribution is wide. None claimed to possess absolute pitch ('perfect pitch').

Forty-four participants were chosen in order to ensure each stimulus (as characterized by its matched and unmatched timbral tunings) was

Table 4.1: Generator sizes (cents) for the different tunings.

3-TET	4-TET	5-TET	7-TET	10-TET	11-TET	12-TET	13-TET	15-TET	16-TET	17-TET
400	300	720	685.71	360	327.27	700	369.23	320	375	705.88

tested by a number of participants sufficiently large to detect small-sized effects and to ensure a broad range of participants took part (as characterized by musical experience, taste, age, etc.). Due to the experimental design, each such stimulus was rated by an average of twenty-four participants.

4.2.2 *Stimuli and Apparatus*

The tones were generated by The Viking (Milne and Prechtl, 2008), a freeware additive-subtractive synthesizer built within Outsim's SynthMaker,[11] which can be downloaded from `http://www.dynamictonality.com`. The synthesizer allows for tones to be played in a wide variety of tunings and, furthermore, for the tones' partials to be matched to those tunings; by adapting the code, I made it possible also to produce unmatched timbres (where the timbre is matched to an n-TET different to the underlying tuning). In a matched timbre, the interval size (in a log-frequency measure like cents) between any possible pair of partials corresponds to an interval found in the underlying tuning—in other words, the partials are themselves tuned to the underlying n-TET. The Viking supports spectral matching to some, but not all, n-TETs. The n-TETs with the eleven lowest values of n, currently supported by The Viking, were those chosen for the experiment (i.e., 3-TET, 4-TET, 5-TET, 7-TET, 10-TET, 11-TET, 12-TET, 13-TET, 15-TET, 16-TET, and 17-TET).

The method used to match the spectral tuning to the underlying tuning is fully described in the Dynamic Tonality section of Sethares et al. (2009).[12] But, in brief, every pitch is expressed as a different linear

11 `http://synthmaker.co.uk`

12 The MATLAB code, which replicates this method for the model, is in the 'Preliminary (run once) calculations - prime mappings etc.' section of the file Affin-

combination of two intervals—a *period* and a *generator*. This means that, given a period and generator of specific sizes, an entire family of pitches is produced. For this experiment, the period was fixed to 1200 cents, which is the standard octave, while the generator takes a different size according to the tuning system required. For example, a generator of 700 cents produces 12-TET, while a generator of 360 cents produces 10-TET. The generator sizes for the eleven tunings used in this experiment are shown in Table 4.1.

In the same way that every fundamental pitch is constructed from a linear combination of the period and generator, so are its partials (in a matched timbre). For example, in a timbre matched to a 10-TET tuning, the second harmonic is equivalent to one period above the fundamental (1200 cents), the third harmonic is equivalent to three periods minus five generators ($3 \times 1200 - 5 \times 360 = 1800$ cents), the fourth harmonic is equivalent to two periods (2400 cents), the fifth harmonic is equivalent to two periods plus one generator ($2 \times 1200 + 360 = 2760$ cents). The precise linear combinations that lead to these approximations of each harmonic are produced by transformation (prime-mapping) matrices that minimize the amount by which the tunings of the partials are altered from pure harmonicity at specific generator sizes (Milne et al., 2008; Sethares et al., 2009). The Viking uses three transformation matrices: one for generators in the range 685.71–720 cents (the syntonic mapping), which includes 7-TET, 12-TET, 17-TET and 5-TET; another for generators in the range 360–400 cents (the magic mapping), which includes 10-TET, 13-TET, 16-TET, and 4-TET; another for generators in the range 300–327.27 cents (the hanson mapping), which includes 4-TET, 15-TET, and 11-TET. The tuning values for the first twelve partials, and that of a harmonic complex tone are shown in Table 4.2 (The Viking's partials extend up to the 32nd (tempered) harmonic, but only the first twelve are considered in the harmonicity and pitch similarity models—see Section 4.3.2.3).

---

ity_data_analysis.m, which can be downloaded from `http://www.dynamictonality.com/melodic_affinity_files/`.

Table 4.2: The tunings (relative to the first partial and rounded to the nearest cent) of the partials of a harmonic complex tone (HCT) and the timbres matched to the n-TETs used in the experiment.

	Partial number											
Spectrum	1	2	3	4	5	6	7	8	9	10	11	12
HCT	0	1200	1902	2400	2786	3102	3369	3600	3804	3986	4151	4302
3-TET	0	1200	2000	2400	2800	3200	3600	3600	4000	4000	4000	4400
4-TET	0	1200	1800	2400	2700	3000	3300	3600	3600	3900	4200	4200
5-TET	0	1200	1920	2400	2880	3120	3600	3600	3840	4080	3840	4320
7-TET	0	1200	1886	2400	2743	3086	3257	3600	3771	3943	4286	4286
10-TET	0	1200	1800	2400	2760	3000	3120	3600	3600	3960	4320	4200
11-TET	0	1200	1964	2400	2836	3164	3382	3600	3927	4036	4145	4364
12-TET	0	1200	1900	2400	2800	3100	3400	3600	3800	4000	4100	4300
13-TET	0	1200	1846	2400	2769	3046	3231	3600	3692	3969	4246	4246
15-TET	0	1200	2000	2400	2800	3200	3600	3600	4000	4000	4000	4400
16-TET	0	1200	1875	2400	2775	3075	3300	3600	3750	3975	4200	4275
17-TET	0	1200	1906	2400	2824	3106	3459	3600	3812	4024	4024	4306

The synthesizer was controlled by Cycling 74's Max/MSP.[13] I created a patch that generated random melodies, each containing 16 eighth-notes (i.e., two bars of 4/4). Stochastic (transition) matrices were used to model some general features found in melodies: (a) in Western and non-Western melodies, smaller intervals typically occur more often than large intervals (Vos and Troost, 1989, and references therein); (b) the average notated pitch of both Western and non-Western music is approximately D♯4 (Parncutt, 1992) as cited by Huron (2001); (c) intervals that span a small number of fifths (e.g., diatonic intervals) are more common than those that span a large number of fifths (e.g., chromatic intervals)—this can be generalized to microtonal tunings by favouring intervals that span fewer generators (the 700 cent fifth being the generator of 12-TET); (d) modulations are infrequent (this is implicit in the very notion of 'scale'—if a scale modulates frequently, it loses its privileged status; for instance, if a diatonic scale tuned to 12-TET is randomly modulated every other note, the resulting scale is more sensibly described as being 12-TET rather than diatonic). I model each of these features with four probability mass functions. But, before describing them, it is necessary to explain how the melodies were generated.

13 http://cycling74.com

Every pitch was characterized by an octave coordinate j and a generator coordinate k. These are the numbers of octaves (periods) and generators above the *reference pitch* D4 (≈ 293.7 Hz—which is close to the average notated pitch D♯4). As discussed above, each different n-TET has a differently sized generator, as shown in Table 4.1. For example, in 12-TET the generator is a 700 cent fifth. This means the pitch D4 has the (j, k) coordinates $(0, 0)$ because it is equivalent to the reference note D4; the pitch A5 has coordinates $(1, 1)$ because it is one octave and one fifth above the reference D4; the pitch A3 has coordinates $(-1, 1)$ because it is down one octave and up one fifth from D4. For 10-TET, the generator is 360 cents. This means that, in this tuning, the coordinates $(0, 0)$ still refer to the reference note D4, but $(1, 1)$ now refer to a pitch one octave (1200 cents) plus one generator (360 cents) above D4 (i.e., 1560 cents above D4); and so forth.

This method of generating pitches also means that every pitch can be reached by more than one set of coordinates. For example, in 12-TET the coordinates $(7, -12)$, $(0, 0)$, and $(-7, 12)$ all point to D4. It also means the cents pitch of any note with coordinates (j, k) can be calculated as $1200j + \beta k$, where β is the size in cents of the generator.

Let me define four variables: *pitch height* is the current pitch (at time t) in cents above D4 (i.e., $1200j_t + \beta k_t$); *pitch change* is the pitch difference between the next pitch (which is at time $t + 1$) and the current pitch (i.e., $1200(j_{t+1} - j_t) + \beta(k_{t+1} - k_t)$); *generator location* is simply the k-coordinate of the current pitch; *generator change* is the difference between next pitch's k_{t+1} coordinate and the current pitch's k coordinate (i.e., $k_{t+1} - k_t$). In the context of well-formed scale theory, pitch classes with small absolute values of k can be thought of as more 'diatonic', while those with large absolute values can be thought of as more 'chromatic'. For example, putting the 12-TET pitch classes in order of fifths (E♭–B♭–F–C–G–D–A–E–B–F♯–C♯–G♯), it is clear the diatonic pitches are closest to the central pitch class D (they have $-3 \leq k \leq 3$), while the chromatic pitches are the most distant (they have $k \leq -4$ and $4 \leq k$). Similarly, intervals with small absolute generator changes can

be thought of as more 'diatonic' and those with large absolute generator changes can be thought of as more 'chromatic'.

The following four probability mass functions were used to model the four common melodic features discussed above:

(a) Smaller intervals were favoured over larger intervals by using a symmetrical triangular probability mass function of *interval size*, centred at zero with a full width at half maximum of 1000 cents.[14] This can be represented by a first-order Markov process, and hence by a stochastic matrix $\mathbf{P}_{\mathrm{RelPitch}} \in [0,1]^{N\times N}$, where N is the number of different pitches.

(b) Pitches in the middle of the musical pitch range were favoured over lower or higher pitches by using a symmetrical triangular probability mass function of *pitch*, centred at zero, with a full width at half maximum of 1200 cents. This can be represented by a zeroth-order Markov process, and hence by a stochastic matrix $\mathbf{P}_{\mathrm{AbsPitch}} \in [0,1]^{N\times N}$, all of whose rows are identical.

(c) 'Diatonic' intervals were favoured over 'chromatic' by using a symmetrical triangular probability mass function of *generator change*, centred at zero with a full width at half maximum of 6.5 generators. These probabilities are then assigned to the pitches arranged in the same as order as the first two matrices. Because more than one set of (j,k) coordinates may produce the same pitch, their probabilities summed. This can be represented by a first-order Markov process, and hence by a stochastic matrix $\mathbf{P}_{\mathrm{RelGen}} \in [0,1]^{N\times N}$.

(d) A central generator location was favoured over outlying locations by using a symmetrical triangular probability mass function of the *generator location*, centred at zero with a full width at half maximum of 6 generators. As before, the probability assigned to each pitch is given by the sum of the above probabilities for all (j,k)

14 For the discrete random variable X, a symmetrical triangular probability mass function, with central value c and full width at half maximum w, is defined accordingly: for $c-w<x<c+w$, $p_X(x) = (1-|x|/w)/\sum_x 1-|x|/w$.

coordinates equivalent to that pitch. This can be represented by a zeroth-order Markov process, and hence by a stochastic matrix $\mathbf{P}_{\text{AbsGen}} \in [0,1]^{N\times N}$, all of whose rows are identical.

The final stochastic matrix of pitch transitions $\mathbf{P}_{\text{Pitch}}$ is given by an appropriately normalized entrywise product of the above four stochastic matrices; that is,

$$\mathbf{P}_{\text{Pitch}} = \mathbf{DM}\text{, where}$$
$$\mathbf{M} = \mathbf{P}_{\text{RelPitch}} \circ \mathbf{P}_{\text{AbsPitch}} \circ \mathbf{P}_{\text{RelGen}} \circ \mathbf{P}_{\text{AbsGen}}\,, \tag{4.1}$$

where $\circ$ denotes the Hadamard (entrywise) product, and $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal row-normalization matrix that ensures the elements in every row of $\mathbf{P}_{\text{Pitch}}$ sum to 1; that is,

$$\mathbf{D} = \begin{pmatrix} |\mathbf{m}_1|^{-1} & 0 & 0 & \dots & 0 \\ 0 & |\mathbf{m}_2|^{-1} & 0 & \dots & 0 \\ 0 & 0 & |\mathbf{m}_3|^{-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & |\mathbf{m}_N|^{-1} \end{pmatrix}, \tag{4.2}$$

where $|\mathbf{m}_n|^{-1}$ is unity divided by the sum of the elements in the nth row of $\mathbf{M}$.

The values, given above, for the means and spreads of the triangular distributions used to create the four stochastic matrices $\mathbf{P}_{\text{RelPitch}}$, $\mathbf{P}_{\text{AbsPitch}}$, $\mathbf{P}_{\text{RelGen}}$, and $\mathbf{P}_{\text{AbsGen}}$, were initially chosen by informed guesswork and then refined, by trial and error, to produce musically pleasing melodies. Furthermore, for each different melody the inter-onset-interval for eighth-notes was randomly chosen, with a uniform distribution, over the range 163–476 ms (63–184 beats per minute), whose mean of 319.5 ms (94 bpm) equates to a medium tempo; the articulation (ratio of note-length to inter-onset-interval) was randomly chosen from the range 0.72 to 0.99, whose mean of 0.86 equates to the average articulation used by organists (Jerket, 2004).

The timbre used was moderately bright and had a quick, but non-percussive sounding, attack and a full sustain level. With harmonic par-

tials, it sounded somewhat like a brass or bowed-string instrument. It was created by using a spectrum with partial amplitudes of $1/i$ where i is the number of the partial (if all partials had been in the same phase and tuned to a harmonic series this would give a sawtooth waveform). To slightly mellow the timbre, the tones were then passed through The Viking's low-pass filter set to give a small resonant peak. A small amount of delayed-onset vibrato was added to give the sound life, and a small amount of reverb/ambience to emulate the sound of a small recital room. The stimuli were listened to with closed-back circumaural headphones in a quiet room. The adapted version of The Viking used in the experiment, the Max/MSP patch that generated the random melodies, and audio files for a sample of the stimuli can be downloaded from `http://www.dynamictonality.com/melodic_affinity_files/`.

4.2.3 *Procedure*

Each participant listened to 60 different randomly generated melodies. Each melody was played in an n-TET randomly chosen from eleven possibilities: 3-TET, 4-TET, 5-TET, 7-TET, 10-TET, 11-TET, 12-TET, 13-TET, 15-TET, 16-TET, and 17-TET. For each melody, the participant could use a mouse or touchpad to switch (toggle) between two different timbres: one timbre was *matched* (its partials matched the underlying tuning); the other was *unmatched* (its partials were matched to a different n-TET randomly chosen from the same list). Each melody could be repeated, by the participant, as many times as wished; most trials were completed in 25–30 minutes. For each participant, no pair of underlying tuning and unmatched spectral tuning occurred more than once. For each pair of stimuli, the participant was asked to make a single choice of timbre for which all or most of the following criteria were best met for the different notes of the melody:

- they have the greatest 'affinity'
- they 'fit together' best

- they sound most 'in-tune' with each other
- they sound the least 'surprising'.

Without further experimental tests, it is impossible to say whether or not these four features are measuring the same latent concept of 'affinity' for all participants. However, for this work, these four descriptions constitute my operationalization of 'affinity'.

The following data were recorded for each melody:

- the tuning of the melody, and the tuning of the unmatched timbre's partials
- whether the matched timbre (coded with a 1) or unmatched timbre (coded with a 0) was chosen
- the tempo and articulation values

For each participant, age, sex, musical taste, and musical experience or training were also collected. General comments were also asked for.

When stimuli are characterized by their underlying tuning (or, equivalently, the tuning of the matched timbre's partials) and the tuning of the unmatched timbre's partials (which must be different, by definition), there are ${}_{11}P_2 = 110$ different possible stimuli. The 60 different stimuli listened to by each participant were sampled randomly without replacement (uniform distribution) from the 110. This means that, on average, each possible pair of underlying tuning and unmatched tuning has been tested $44 \times 60/110 = 24$ times, each underlying tuning (row of Fig. 4.6) $44 \times 60/10 = 264$ times, each unmatched spectral tuning (column of Fig. 4.6) $44 \times 60/10 = 264$ times. In total there were $44 \times 60 = 2640$ observations of 110 different stimuli.

4.3 RESULTS

The experimental data, aggregated over all participants, are summarized in Figure 4.6. The squares represent the 110 different pairs of stimuli the participants were presented with; the shade of each square

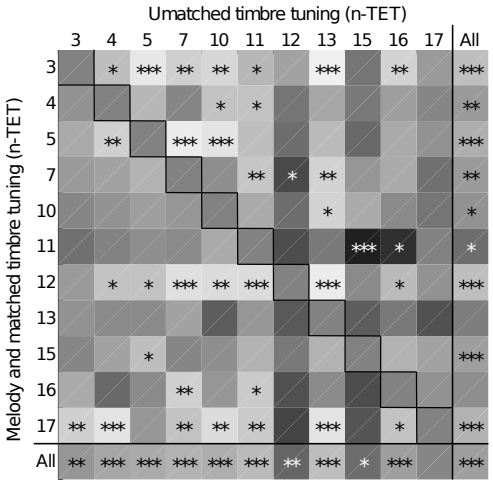
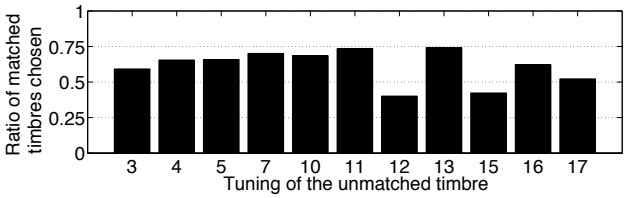


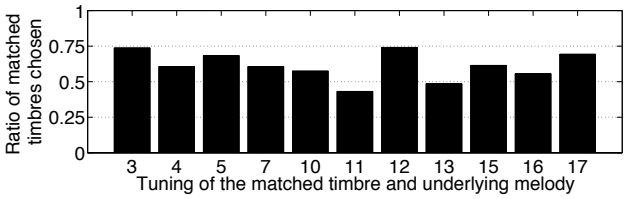
Figure 4.6: Results aggregated over all participants: the squares represent the 110 different pairs of stimuli the participants were presented with; the shading of each square indicates the ratio of occasions when the matched, rather than unmatched, timbre was chosen (white for all matched, black for all unmatched). The rows (labeled by n-TET) represent the different underlying tunings (or, equivalently, the matched timbres' spectral tuning); the columns represent the different unmatched timbres' spectral tunings. The bottom row and rightmost column show ratios aggregated over underlying tunings and unmatched timbres, respectively. The bottom-right square is the ratio aggregated over all tunings and unmatched timbres. Black stars indicate significantly more than half of the choices were for matched timbres, white stars indicate significantly more were for unmatched timbres using a two-tailed exact binomial test (* for $p < .05$, ** for $p < .01$, *** for $p < .001$). Bonferroni correction has been not been applied.

indicates the ratio of occasions when the matched timbre was chosen over the unmatched. More precisely: the vertical axis shows the n-TET used for the underlying tuning (which, by definition, is equivalent to the tuning of the matched timbres' partials); the horizontal axis shows the n-TET used for the tuning of the unmatched timbres' partials; the grey-scale indicates the probability of the matched timbre being chosen (black corresponds to a probability of zero, white to a probability of one). For example, the square on the row marked 7 and the column marked 11, shows the proportion of occasions that, for a 7-TET melody, a matched timbre (partials tuned to 7-TET) was chosen in preference to an unmatched timbre with partials tuned to 11-TET.

The squares in the top-left to bottom-right diagonal (they have thicker borders) correspond to situations where both stimuli are iden-



(a) Over different unmatched timbres—the bottom row of Fig. 4.6.



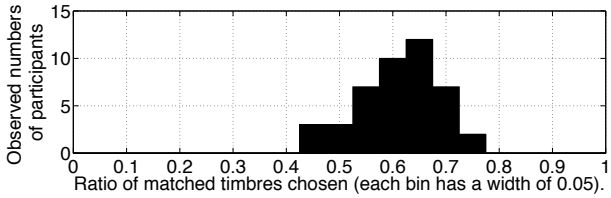
(b) Over different matched timbres (underlying tunings)—the rightmost column of Fig. 4.6.

Figure 4.7: Ratios of matched timbres chosen.

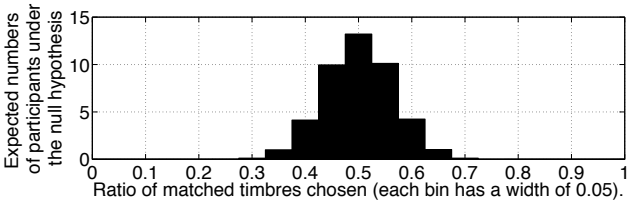
tical (they are both equally matched to the underlying tuning). Such pairs were not tested because it is clear that—given the forced-choice nature of the procedure—the probability of choosing either stimulus would converge to .5. For this reason, the diagonal is shaded accordingly, and this serves as a useful reference against which to compare the other data points.

The bottom row shows the ratio of occasions a matched timbre was chosen, aggregated over all possible tunings, for each of the eleven unmatched timbres (this is also shown in Fig 4.7a). The rightmost column shows the ratio of occasions a matched timbre was chosen, aggregated over all possible unmatched timbres, for each of the eleven underlying tunings (this is also shown in Fig. 4.7b). The bottom-right square shows the ratio of occasions a matched timbre was chosen aggregated over all underlying tunings and unmatched timbres.

A single star indicates a proportion that is significantly different from .5 (using a two-tailed exact binomial test) at a level of .05, two stars indicate significance at the .01 level, three stars at the .001 level. I have not applied Bonferroni correction, because I am not inferring a preference for matched partials on the basis of any single stimulus, and it is interesting to see which of the stimuli are sufficiently different from chance to merit individual significance (it is worth noting that with 110 sep-



(a) As observed in the experiment.



(b) As expected under the null hypothesis.

Figure 4.8: Histograms of participants' ratios of matched timbres chosen—as observed in the experiment (a), and as expected under the null hypothesis that pitch similarity has no association with affinity (b). The values for (b) are the means of multiple histograms randomly generated under the null hypothesis.

arate tests we would expect 5.5 to be significant at the .05 level under the null hypothesis of pure chance).

Figure 4.8a is a histogram showing the distribution of participants' responses aggregated over all stimuli. Their responses are measured by the ratio of matched timbres chosen, and placed into bins of width 0.05.[15] For comparison, Figure 4.8b shows the histogram that would be expected under the null hypothesis that pitch similarity has no impact on affinity (i.e., the probability of choosing a matched timbre is .5 for all stimuli and all participants).

Under the null hypothesis that matched and unmatched timbres are chosen with equal probability, all squares in Figure 4.6 would have a fairly uniform mid-grey appearance, all the bars in Figure 4.7 would be close to .5, and the histogram of participants' responses would be centred at .5 (as in Figure 4.8b). The data appear to show two clear aspects that differ from the null hypothesis. Firstly, as discussed in Section 4.3.1, the overall ratio of matched timbre choices is significantly greater than chance. In Figure 4.6, this is indicated by the bottom-right

15 This bin width was chosen because the centre values are easy to understand, it divides the overall range of 0 to 1 into 20 bins, and because the number of bins (across the range of the data) follows Sturges' rule (1926) of $\lceil 1 + \log_2 n \rceil$, where n is the number of data points (for this experiment, Sturges rule suggests $\lceil 1 + \log_2 44 \rceil = 7$ such bins).

square and the large number of stimuli where matched timbres were chosen significantly more often than chance. In Figures 4.7 and 4.8a, it is indicated by the bias above .5.

Secondly (as modelled in Sec. 4.3.2), there are some obvious vertical and horizontal stripes in Figure 4.6. For example, the columns representing the unmatched timbres tuned to 12-TET and 15-TET (and possibly 17-TET) are darker (and their aggregated values are relatively low, as shown in Fig. 4.7a). This suggests that the participants felt these timbres tended to have relatively higher affinity regardless of the underlying tuning. This is interesting because these three timbres all have partials that are relatively close to perfectly harmonic partials (our harmonicity model—described in Sec. 4.3.2.3—confirms this). In other words, it suggests a general preference for timbres with high harmonicity. The horizontal stripes, which represent the underlying tuning and its matched timbre, are complementary to the vertical stripes. For example, if 12-TET timbres are preferred regardless of tuning then, when the underlying tuning—and its matched timbre—is 12-TET, more of the unmatched timbres are now less likely to be chosen. Hence, the corresponding row is lighter. So the dark vertical stripes and corresponding light horizontal stripes are complementary manifestations of the same process.

The heavy left-hand tail in Figure 4.8a may indicate the presence of a few participants for whom the impact of pitch distance was negligible. However, without more data, this is impossible to ascertain with certainty.

4.3.1 *Data Aggregated Over All Stimuli and Participants*

The data, aggregated over all possible pairs of stimuli and all participants, are represented by the bottom right square of Figure 4.6. A total of 2,638 different tests, of whether the matched or unmatched timbre was chosen, were performed (44 participants each listening to 60 stimuli, with two tests lost due to the experiment ending prematurely).

In total, matched timbres were chosen 1,615 times (61% of occasions). Given the null hypothesis (that listeners do not think matched timbres give better-fitting, more in-tune, melodies than unmatched), the expected number of matched timbres chosen would be $.5 \times 2638 = 1319$ with a binomial distribution of $\mathrm{Bin}(2638, .5)$. Under this null hypothesis, a two-tailed exact binomial test shows the probability of 1,615, or greater, matched timbres being chosen is $p < .001$.[16] This indicates that affinity is increased with the use of matched timbres (whose partials match the underlying equal tuning) in comparison to unmatched timbres (whose partials match a different equal tuning) and, hence, supports the hypothesis that affinity is a monotonic function of spectral pitch similarity.

Of the 44 participants, 38 (86%) chose matched timbres for more than half of the 60 stimuli they listened to. Under the null hypothesis that 50% of participants would choose matched timbres more often than unmatched, an exact binomial test (two-tailed) shows the probability of this occurring by chance is $p < .001$, thus showing the preference for matched timbres was not confined to a small number of 'high performing' participants, thereby providing further evidence in support of the above hypothesis, and the generality of its conclusions.

In this section, I have restricted myself to analyses of the data aggregated over all tunings and unmatched timbres. It is clear, however, that there is a wealth of information contained in the remainder of the data (such as the stripes discussed above). In the following subsection, I utilize the model illustrated in Figure 4.5 to explore whether it is possible to explain some of the finer patterns of the data illustrated above. I do not, however, test any subject-effects models (i.e., models containing explanatory variables such as participants' musical taste and level of experience). The reason for this is because the resulting models would be too computationally complex to make cross-validation feasible (due to the nonlinear nature of my models, cross-validation is a useful method

16 Indeed, 1370 (52%) is the minimum number of matched timbre choices that would have been significant at the .05 level.

to see how well the models can generalize and to test their statistical significance—as discussed in Sec. 4.3.2.4).

4.3.2 *Data Aggregated Over Participants Only*

In the following three sections (4.3.2.1–4.3.2.3), I describe three submodels: the first based upon spectral pitch similarity, the second on virtual pitch similarity, the third on harmonicity. In Section 4.3.2.4, I show how these submodels are combined, in different ways, to create three candidate models of the experimental data. These three models are used to predict the 110 data values illustrated in Figure 4.6 (every tested pair of matched and unmatched spectral tuning).

As discussed in Section 4.1, I model participants' affinity responses with a combination of spectral pitch similarity (or virtual pitch similarity) and harmonicity. Because I am modelling forced choices made between two alternative stimuli, I am concerned with the differences between the spectral (or virtual) pitch distances, and harmonicities, of the two stimuli in each pair. For spectral and virtual pitch similarity, it is also necessary to take into account that the stimuli are not single intervals, but melodies containing a stream of intervals with varying probabilities of occurring.

The following descriptions are principally verbal, interspersed with necessary mathematical equations; a more formal mathematical description of the model is provided in Appendix D, the relevant equations of which are referenced below. The models and data can be downloaded, as MATLAB .m files, from `http://www.dynamictonality.com/melodic_affinity_files/`.

4.3.2.1 *Spectral pitch similarity comparison submodel*

Given the stochastic parameters described in Section 4.2.2 that were used to generate the random melodies (and which are constant across all stimuli), each stimulus can be characterized by the tuple (m, n_1, n_2), where $m = 1, 2, \ldots, 11$ indexes the underlying tuning, $n_1 = 1, 2, \ldots, 11$

indexes the spectral tuning of the matched timbre, and $n_2 = 1, 2, \ldots, 11$ indexes the spectral tuning of the unmatched timbre (this implies that, for this experiment, $m = n_1$).

Given a melody randomly generated in tuning m, the *spectral pitch similarity submodel* $f_{\mathrm{S}}(m, n_1, n_2; \rho, \sigma)$ is the expected spectral pitch similarity of consecutive tones with matched timbre n_1 minus the expected spectral pitch similarity of consecutive tones with unmatched timbre n_2. As explained below, it is parameterized by ρ and σ.

The first twelve partials of a tone (as shown in Table 4.2) are represented by a spectral pitch vector (as defined in Section 4.1.1) with the ith partial given a salience of $i^{-\rho}$, where ρ is a free parameter representing *spectral roll-off* (D.1–D.2).[17] When $\rho = 0$, all partials have a salience of one; as ρ increases above 0, the saliences of the higher partials reduce. This parameter, therefore, provides a simple way to change the influence of lower versus higher partials.

To model perceptual pitch inaccuracy (the extent to which partials with similar frequencies are heard to have the same pitch), each partial in the spectral pitch vector is smeared in the log-frequency domain by convolving with a discrete and truncated approximation of a normal probability distribution with a standard deviation of σ_{S}, which is a free parameter representing *smoothing width* (D.3).

For all underlying tunings and timbres (matched or unmatched), the spectral pitch similarities of all possible intervals are calculated by taking the cosine similarity of their spectral pitch vectors (D.6). Each melody is generated according to the probability distribution described in Section 4.2.2. This enables us to calculate the *expected spectral pitch similarity* for consecutive tones for any combination of underlying tuning and timbre (values of m and n); this is achieved by multiplying the spectral pitch similarity of each interval by its probability of occurring and summing over all interval sizes (D.7). The reason for using probabilities derived from the Markov matrices that generated the melodies, rather than using the intervals that actually occurred in each and ev-

17 Only the first twelve partials are used because experiments have demonstrated that higher partials are typically not resolvable by the auditory system (e.g., Bernstein and Oxenham (2003)).

ery melody, is to keep the modelling computationally tractable (as it is, each model takes 10 hours to cross-validate—see Sec. 4.3.2.4).

The value of $f_{\mathrm{S}}(m, n_1, n_2; \rho, \sigma)$ is given by subtracting the expected spectral pitch similarity of consecutive tones with timbre n_2 from the expected spectral pitch similarity of consecutive tones with timbre n_1 (D.10). It contains just two free parameters: spectral roll-off $\rho \in (-\infty, \infty)$, and smoothing width $\sigma \in [0, \infty)$. This submodel is used—in part—to model the probability of choosing a matched timbre over an unmatched timbre. For concision, this submodel may be denoted by the abbreviated form f_{S}, and the vector of its values over the 110 tested stimuli is denoted $\mathbf{f}_{\mathrm{S}}$.

4.3.2.2 *Virtual pitch similarity comparison submodel*

Given a melody randomly generated in tuning m (as described in Section 4.2.2), the *virtual pitch similarity submodel* $f_{\mathrm{V}}(m, n_1, n_2; \rho, \sigma)$ is the expected virtual pitch similarity of consecutive tones with matched timbre n_1 minus the expected virtual pitch similarity of consecutive tones with unmatched timbre n_2.

The first twelve partials of the tone and a harmonic template are embedded in spectral pitch vectors, both with smoothing widths of σ cents, and the salience of each ith partial given by $i^{-\rho}$ (D.1–D.3). As described in Section 4.1.1, the spectral pitch vectors for the timbre and the harmonic template are cross-correlated (non-circular cross-correlation is used), to produce a virtual pitch vector that gives the unnormalized correlation between the two vectors over differing offsets between their lowest partials (D.4). This serves as a model for virtual pitch weights at log-frequencies relative to the timbre's lowest partial (this is illustrated in Figure 4.3, which shows the virtual pitch vector for a complex harmonic tone with $\sigma = 10.3$ and $\rho = 0.42$).

For all underlying tunings and timbres (matched or unmatched), the virtual pitch similarities between all possible intervals are calculated by taking the cosine similarities of their virtual pitch vectors and, as

described above, their expected values for all pairs of matched and unmatched timbres are calculated (D.8–D.9).

The value of $f_{\mathrm{V}}(m, n_1, n_2; \rho, \sigma)$ is given by subtracting the expected virtual pitch similarity of consecutive tones with timbre n_2 from the expected virtual pitch similarity of consecutive tones with timbre n_1 (D.11). It contains just two free parameters: spectral roll-off $\rho \in (-\infty, \infty)$, and smoothing width $\sigma \in [0, \infty)$. This submodel is used—in part—to model the probability of choosing a matched timbre over an unmatched timbre. For concision, this submodel may be denoted by the abbreviated form f_{V}, and the vector of its values over the 110 tested stimuli is denoted $\mathbf{f}_{\mathrm{V}}$.

4.3.2.3 *Harmonicity comparison submodel*

The toneness and vertical familiarity of a timbre are both modelled by its harmonicity. The *harmonicity submodel* $f_{\mathrm{H}}(n_1, n_2; \rho, \sigma)$ calculates the harmonicity of the matched timbre minus the harmonicity of the unmatched timbre. It is not affected by the underlying tuning m.

Harmonicity is here calculated as the maximum value in the cross-correlation of the timbre and a harmonic template; that is, it is the maximum value found in the virtual pitch vector (D.5). This means that the more similar the pattern of the timbre's spectral contents is to the harmonic template, the greater its harmonicity.

The value of $f_{\mathrm{H}}(n_1, n_2; \rho, \sigma)$ is given by subtracting the harmonicity of timbre n_2 from the harmonicity of timbre n_1 (D.12). As with the other submodels, it contains just two free parameters: spectral roll-off $\rho \in (-\infty, \infty)$, and smoothing width $\sigma \in [0, \infty)$. They are, however, not necessarily expected to be identical to the related parameters in the spectral pitch similarity model, because it is possible non-identical processes are occurring. In the former, I am modelling a comparison between two events held in short-term memory; in the latter, one of the items—the harmonic template—does not reside in short-term memory (it either resides in long-term memory or is embodied in some innate cognitive process). We would, however, expect the parameters

to be identical to those used in the virtual pitch similarity model, because they both involve comparison with a harmonic template. This submodel is used—in part—to model the probability of choosing a matched timbre over an unmatched timbre. For concision, this submodel may be denoted by the abbreviated form f_{H}, and the vector of its values over the 110 tested stimuli is denoted $\mathbf{f}_{\mathrm{H}}$.

4.3.2.4 *Three candidate models of participants' responses*

The spectral and virtual pitch similarity submodels are for mental processes triggered by successive tones (horizontal musical features); the harmonicity model is for mental processes triggered by individual tones (vertical musical features). I test three different models that combine one of the two horizontal submodels with the vertical submodel (two of the resulting models contain the same two submodels but utilize a different parameterization). In addition to the smoothing width and roll-off parameters required by each submodel, each candidate model requires two additional parameters (β_1 and β_2) to set the relative weights of its two submodels.

The first candidate model utilizes the two submodels spectral pitch similarity $f_{\mathrm{S}}(m, n_1, n_2; \rho_{\mathrm{S}}, \sigma_{\mathrm{S}})$ and harmonicity $f_{\mathrm{H}}(n_1, n_2; \rho_{\mathrm{H}}, \sigma_{\mathrm{H}})$. This model has a total of six parameters: the submodel weights β_1 and β_2, and the independent smoothing widths and roll-offs for both submodels (as shown after the semicolons).

The second candidate model utilizes the same two submodels, but uses identical smoothing widths and roll-offs for both submodels: spectral pitch similarity $f_{\mathrm{S}}(m, n_1, n_2; \rho, \sigma)$ and harmonicity $f_{\mathrm{H}}(n_1, n_2; \rho, \sigma)$. This model, therefore, has a total of four parameters: the submodel weights β_1 and β_2, and the smoothing width and roll-off parameters used for both submodels. It is, therefore, a more parsimonious version of the first model and reflects the possibility that spectral pitch and harmonicity do indeed derive from closely related perceptual processes (see the last paragraph in Section 4.3.2.3).

The third candidate model utilizes the two submodels virtual pitch similarity $f_{\mathrm{V}}(m, n_1, n_2; \rho, \sigma)$ and harmonicity $f_{\mathrm{H}}(n_1, n_2; \rho, \sigma)$. This model, therefore, has a total of four parameters: the submodel weights β_1 and β_2, and the smoothing width and roll-off parameters used for both submodels.

A model containing both $f_{\mathrm{S}}(m, n_1, n_2; \rho, \sigma)$ and $f_{\mathrm{V}}(m, n_1, n_2; \rho, \sigma)$ (i.e., spectral and virtual similarities) was not tested because these two submodels are too highly correlated—$\mathrm{f_S}$ as optimized in the first model and $\mathrm{f_V}$ as optimized in the third model have a correlation over the stimuli of $r(108) = .95$.

The method by which the two submodels are combined is determined by the data—the collected data are the numbers of matched (rather than unmatched) timbres chosen, and the total number of trials for each of the 110 different pairs of stimuli (*observations*). As such, the data for each stimulus are presumed to be random observations from a binomial distribution and are, therefore, most appropriately modelled with a logistic regression upon the submodels ($\mathrm{f_S}$ and $\mathrm{f_H}$ in the first and second models, $\mathrm{f_V}$ and $\mathrm{f_H}$ in the third model).[18] The three candidate models, indexed by i, of the probability of choosing the matched timbre n_1 (coded 1), given a microtonal melody in tuning m, a matched timbre n_1, and an unmatched timbre n_2, take the form

$$\begin{aligned}
\text{model } i &= P\Big(Y = 1 \mid m, n_1, n_2; i\Big) \\
&= \frac{1}{1 + \mathrm{e}^{-z_i}}\text{, where} \\
z_1 &= \beta_1 f_{\mathrm{S}}(m, n_1, n_2; \rho_{\mathrm{S}}, \sigma_{\mathrm{S}}) + \beta_2 f_{\mathrm{H}}(n_1, n_2; \rho_{\mathrm{H}}, \sigma_{\mathrm{H}}) \\
z_2 &= \beta_1 f_{\mathrm{S}}(m, n_1, n_2; \rho, \sigma) + \beta_2 f_{\mathrm{H}}(n_1, n_2; \rho, \sigma) \\
z_3 &= \beta_1 f_{\mathrm{V}}(m, n_1, n_2; \rho, \sigma) + \beta_2 f_{\mathrm{H}}(n_1, n_2; \rho, \sigma)\,, \qquad (4.3)
\end{aligned}$$

and β_1 and β_2 are the logistic parameters (coefficients). Note that a constant term is not used in this logistic regression. This is because $f_{\mathrm{S}} = f_{\mathrm{V}} = f_{\mathrm{H}} = 0$ when $n_1 = n_2$, which results in a predicted prob-

18 The use of a binomial distribution implies that participants all had the same probability of choosing a matched timbre for each stimulus. Clearly, this is a simplification.

Table 4.3: The log-likelihoods $\log(L)$ for the three models (higher is better).

	Model 1	Model 2	Model 3
$\log(L)$	-257.60	-257.76	-258.84

ability of $.5$ for the matched (or unmatched) stimulus being chosen. In this circumstance, both timbres are identical (they are equally matched to the underlying tuning, and both have the same harmonicity) so the model should predict a .5 probability. A constant term would, therefore, be counterintuitive (to test this, model 2 with an additional constant term was optimized and the term was, as expected, very close to zero and not significant.)

Although the three models in (4.3) superficially appear to be generalized linear (logistic) regression models, it is important to note that they are actually fully nonlinear. This is because the predictors (f_{S}, f_{V}, and f_{H}) are nonlinear with respect to the parameters ρ and σ, and these parameters are optimized simultaneously with the logistic weights β_1 and β_2. This nonlinearity means there is no simple way to calculate the degrees of freedom of the models (or their predictors), so the standard χ^2 significance tests used for logistic regression models are not appropriate. For this reason, the significance of each model (and each of the individual submodels) was tested by five runs of a 10-fold cross-validation.

The models' parameter values were iteratively optimized in MATLAB using the `fmincon` routine. The parameters were optimized to maximize the likelihood of the model given the data (under the presumption that the numbers of matched timbres chosen are binomially distributed). (As a nonlinear optimization, the resulting parameter values may produce a local, not the global, likelihood maximum). A plot of the deviation residuals against predicted values confirmed the errors were randomly distributed with no apparent pattern.

After optimization, the likelihoods of the three models were similar, as shown in Table 4.3. To assess how well the models are able to generalize, and to test their statistical significance (as well as the sig-

nificance of each predictor), they were subjected to five runs of 10-fold cross-validation. In a 10-fold cross-validation, the empirical data is partitioned into 10 randomly chosen equally-sized folds. One of the partitions is removed and designated the *validation set*, while the remaining nine tenths of the data are designated the *training set*. A model's parameters are optimized (by maximum likelihood) to best fit the training set, and this model is tested (scored) against the validation set. After this, a different partition is chosen for the validation set, and the process is repeated. This is done ten times (thereby ensuring all ten partitions have been used as validation sets) and the scores resulting from each validation are combined to produce a complete set of scores. This comprises a single 10-fold cross-validation.

The resulting scores are subject to an unknown variance, so it advisable to minimize the variance by repeating the 10-fold cross-validation as many times as is practicable (each time with a different partitioning), and taking the means of the scores across the different cross-validation runs (e.g., Kim (2009)). Five runs of the 10-fold cross-validation took approximately 10 hours for each model tested.

I scored each validation with the quadratic proper scoring rule, which measures the accuracy of probabilistic predictions of discrete and mutually exclusive outcomes. For binary data, as elicited in my experiment, the score is calculated accordingly:

$$Q(\boldsymbol{r}, i) = 2r_i - r_i^2 - (1 - r_i)^2 \,, \tag{4.4}$$

where r_i is the probability assigned by the model to the response made by the participant. This gives a score for an average of approximately 264 observations in each validation set (11 stimuli each rated by an average of 24 participants). The scores were averaged across participants, resulting in a mean score for each stimulus; hence 11 mean scores per validation set, and 110 mean scores across the 10-fold validation. The 10-fold cross-validation was performed five times, and the means of the 110 scores across the five runs were taken.

Table 4.4: For the three candidate models and two of the submodels (the submodel denoted $G(f_{\mathrm{S}})$ is a logistic regression with just spectral pitch distance, and $G(f_{\mathrm{H}})$ is a logistic regression with just harmonicity), this table shows the medians of the means of the quadratic scores obtained over five runs of a 10-fold cross-validation. See the main text for a precise description of how these values are calculated. Model 2 has the highest median score. All scores are significantly higher than the null model ($p < .001$) under a Wilcoxon signed-rank test.

Null	Model 2	$G(f_{\mathrm{S}})$	$G(f_{\mathrm{H}})$	Model 1	Model 3
0.500	0.553	0.525	0.522	0.552	0.552

In order to test the significance of each model, I calculated the differences between its 110 mean scores and those produced by a *null model* with no predictors (the null model gives a probability of $.5$ to the outcome being either the 'matched' or 'unmatched' timbre—it represents, therefore, 'pure guesswork'). I tested the resulting 110 values with the Wilcoxon signed-rank test to determine whether the medians of the tested and null models' mean scores were significantly different. This information is summarized in Tables 4.4 and 4.5. The tests indicate that model 2 generalizes better than model 1 (this suggests that the additional flexibility in model 1, which arises from its two additional parameters, seems to be fitting the noise in the data rather than the underlying distribution). There is no significant difference in the scores given to models 2 and 3. However, due to model 2 getting a higher score in the cross-validation, a higher likelihood across all the data, and being computationally simpler, I focus my attention on this model for the remainder of this chapter. This should not, however, be taken as strong evidence that the underlying mental process is down to spectral rather than virtual pitches; furthermore, a different model of virtual pitch similarity might have performed and generalized better.

Tables 4.4 and 4.5 also provide statistics for cross-validations applied to a logistic regression with just spectral pitch distance (denoted $G(f_{\mathrm{S}})$), and a logistic regression with just harmonicity (denoted $G(f_{\mathrm{H}})$). The values of ρ and σ were free in both submodels, not fixed to the optimal values for model 2. In both cases, each of these submodels pro-

Table 4.5: The z-scores and significance levels resulting from Wilcoxon signed-rank tests on the differences between the score medians (given in Table 4.4) of the model identified in the top header and the model identified in the left column. The null model is one with no predictors (not even a constant term) and represents assigning a probability of .5 to the matched or matched timbre being chosen. The label *** indicates significance at the .001 level. The first column demonstrates that model 2 is significantly better than all the other models except model 3 (it is better, but not significantly so). The first row shows that each of the two submodels (spectral pitch distance and harmonicity) is significantly better than the null model, while the first column shows that both submodels together (i.e., model 2) is better than either submodel alone.

	Model 2	$G(f_{\mathrm{S}})$	$G(f_{\mathrm{H}})$
Null	-9.10^{***}	-6.35^{***}	-6.73^{***}
$G(f_{\mathrm{H}})$	-8.75^{***}		
$G(f_{\mathrm{S}})$	-8.55^{***}		
Model 1	-4.01^{***}		
Model 3	-1.14		

duces a significantly higher score than the null model, while model 2—which uses both of them—has a significantly higher score than either of the submodels alone. The effect sizes of the spectral pitch distance and harmonicity models are similar, but their effects are different and hence complementary. This is confirmed by the significantly higher score achieved by model 2, which contains both spectral pitch similarity and harmonicity.

For model 2, the optimized parameter values and their standard errors, and statistical tests for the whole model are summarized in Table 4.6. The standard errors were calculated from a numerically estimated Hessian matrix of the optimized parameter values. The Wilcoxon signed-rank test for the whole model, as detailed above, is highly significant ($z = -9.10, p < .001$), and a Hosmer-Lemeshow test indicates the model's predictions are not significantly different to the observed data ($\chi^2(8, N = 2638) = 4.60, p = .800$). With the optimized values for ρ and σ, the two predictors $\mathrm{f_S}$ and $\mathrm{f_H}$ have low correlation over all stimuli ($r(108) = -.09, p = .334$), so there are no concerns with multicollinearity. Figure 4.9 is a scatter plot, for all 110 stimuli, of

Table 4.6: Statistical analysis and evaluations of the model and its parameters (the logistic part of the model does not include a constant term). Standard errors were derived from a numerical estimation of the Hessian matrix, the z-score and p-value were calculated from a signed-rank test on the cross-validation as described in the main text.

Parameter	Value	SE	95% CI	
σ (smoothing)	10.28	0.70	8.92	11.65
ρ (roll-off)	0.42	0.11	0.21	0.63
β_1 (spectral similarity weight)	3.55	0.51	2.55	4.55
β_2 (harmonicity weight)	44.00	16.34	11.97	76.04
Overall model				
Wilcoxon signed-rank: $z = -9.10$, $p < .001$				
Hosmer-Lemeshow: $\chi^2(8, N = 2638) = 4.60$, $p = .800$				

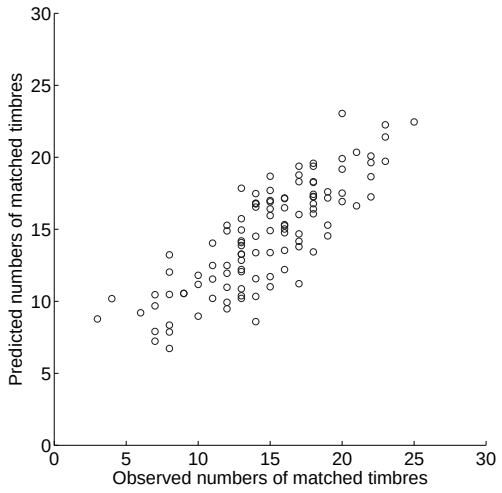


Figure 4.9: For all 110 observations, this scatter plot compares the observed numbers of matched timbres chosen by participants with those predicted by model 2.

the observed numbers of matched timbre choices against the predicted number of matched timbre choices.

For those more accustomed to linear regression statistics, I provide correlation, standardized coefficient, and R^2 values from a linear regression of the log-odds (logits) of choosing a matched timbre. The values of ρ and σ were fixed to the optimized values (0.42 and 10.28)

obtained for model 2, and the following 'linear' model with a constant term was optimized in the standard way with ordinary least squares:

$$\log\left(\frac{P(Y=1 \mid m, n_1, n_2)}{1-P(Y=1 \mid m, n_1, n_2)}\right) = \beta_0 + \beta_1 f_{\mathrm{S}}(m, n_1, n_2; 0.42, 10.28) + \beta_2 f_{\mathrm{H}}(n_1, n_2; 0.42, 10.28)\,. \tag{4.5}$$

Because the values of $\rho = 0.42$ and $\sigma = 10.28$ were originally optimized in model 2, and f_{S} and f_{H} are nonlinear with respect to them, it would be misleading to provide standard F-statistics and F-tests (the degrees of freedom cannot be established), so these are not reported. The optimized model has $R^2 = .59$—that is, 59% of the variance in the log-odds of choosing a matched timbre is accounted for by this 'linear' version of the model. The standardized coefficient values are .38 for spectral similarity $\mathrm{f_S}$ and .70 for harmonicity $\mathrm{f_H}$ (the unstandardized coefficients are 3.54, 47.47, and .05 for the intercept term—similar to those of the logistic model). The correlation between the log-odds of choosing a matched timbre and $\mathrm{f_S}$ over all stimuli is $r(108) = .32, p < .001$, the correlation between the log-odds and $\mathrm{f_H}$ is $r(108) = .66, p < .001$. Following Cohen's familiar guidelines on categorizing effect sizes, spectral similarities have a medium effect size, while harmonicity—and the complete model—have large effect sizes.

The optimized parameter values for the smoothing width and spectral roll-off (10.28 cents and 0.42, respectively) are reassuringly plausible. Under experimental conditions, the frequency difference limen (just noticeable difference) corresponds to approximately 3 cents, which would be modelled by a smoothing width of 3 cents (as explained in App. A). In an experiment like this, in which the stimuli are more explicitly musical, we would expect the standard deviation to be somewhat wider than this, and the value of approximately 10 cents seems eminently reasonable.

The roll-off values are also highly plausible. The optimized roll-off in saliences (0.42) approximately corresponds to the loudnesses of the

partials in the stimuli's timbres.[19] This value emphasizes the importance of the lower partials, and suggests correspondences between the tempered third partial of the timbre and the tempered perfect fifths and fourths in the melody are important (when using matched timbres, tempered perfect fifths and fourths have frequency ratios of tempered 3/2 and tempered 4/3, respectively).

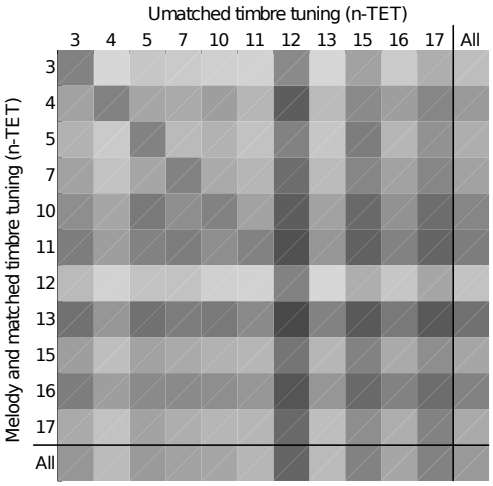
The data generated by the second model are shown in Figure 4.10a, and can be usefully compared with the observed data shown in Figure 4.6. The individual contributions of spectral pitch similarity and harmonicity predictors are shown in Figures 4.10b and 4.10c (in both cases the parameter values are identical to those used in the full model).

4.4 DISCUSSION

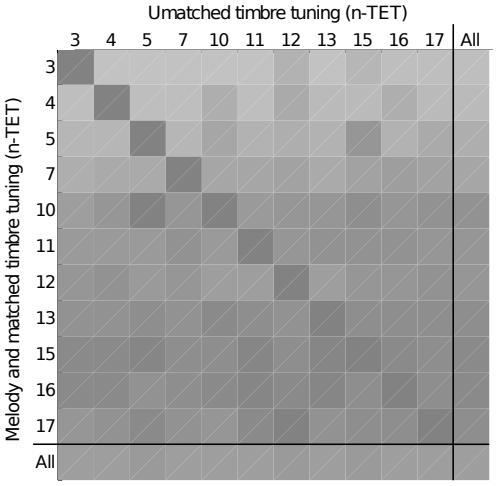
The experimental data strongly support the hypotheses that melodic affinity is increased when the tunings of scale degrees and partials are matched, and when the tones have close-to-harmonic partials. I have also shown how a combination of spectral pitch similarity and harmonicity can model, in a more precise way, the relationship between spectral tuning, melodic tuning, and perceived affinity. Importantly, the experimental procedure allows us to eliminate the confounding top-down influence of horizontal familiarity (that part of affinity that is a function of each interval's prevalence). In the absence of this confound, we can see that spectral pitch similarity has a medium-sized effect on perceived affinity. As explained in Section 4.1.2, the experiment cannot determine whether the impact of harmonicity is because it models our familiarity with harmonic complex tones, or whether it is modelling an innate process—but, either way, its effect size is strong.

It is interesting to explore the implications of these results in two different areas: the use of microtonally tempered spectra for microtonal scales, and the relationship between scales and tunings that pro-

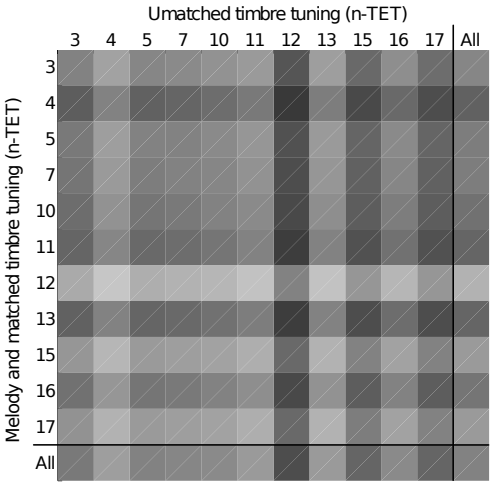
19 The timbres had partials with amplitudes of approximately i^{-1}, where i is the partial number. According to Steven's power law, perceived loudness corresponds, approximately, to amplitude (pressure) to the power of 0.6, hence the loudness of each partial is approximately $i^{-0.6}$, which would be equivalent to a ρ value of 0.6.



(a) Data simulated by model 2—logistic regression on spectral pitch similarity $f_{\mathrm{S}}(m, n)$ and harmonicity $f_{\mathrm{H}}(m, n)$.



(b) Data simulated by a logistic regression on spectral pitch similarity $f_{\mathrm{S}}(m, n)$.



(c) Data simulated by a logistic regression on harmonicity $f_{\mathrm{H}}(m, n)$.

Figure 4.10: The modelled data.

vide good-fitting melodies when using tones produced by conventional wind and string instruments and the human voice (the majority of which have close-to-harmonic partials).

4.4.1 *Matching Timbre and Tuning*

In the Dynamic Tonality section of Sethares et al. (2009), a procedure is given to retune partials to match a wide variety of scale-tunings (scales generated by two intervals—a period and a generator—both of which can take any size). The procedure was aesthetically motivated on the grounds that it can reduce the sensory dissonance of prevalent intervals and chords in the underlying scale. The experimental data, presented here, suggest the matching of partials to (low-n) n-TETs can also make microtonal melodies more in-tune and fitting. Indeed, it was my practical experience with Dynamic Tonality synthesizers—noticing, for example, how much more in-tune 5-TET melodies sound when the spectral tuning is matched—that motivated this experiment in the first place.[20] Having said that, it is also clear that timbres with partials close in frequency to the familiar harmonic template were typically preferred by participants. This means that, in matching partials to a low n n-TET, one is trading the increased consonance and affinity of intervals for possibly dissonant timbres.

4.4.2 *Scales for Tones with Harmonic Spectra*

The majority of pitched Western instruments have timbres whose partials are tuned to a harmonic series (e.g. bowed string, wind instruments, and the voice), or close to such a spectrum (e.g., plucked and hammered string instruments). Figure 4.11 shows the spectral pitch similarity of pairs of tones with harmonic spectra separated by an inter-

20 There are, currently, three Dynamic Tonality synthesizers (which allow for a variety of tunings and for the spectral tuning to be matched)—TransFormSynth (an analysis-resynthesis synthesizer),The Viking (an additive-subtractive synthesizer), and 2032 (a modal physical modelling synthesizer). They are freeware, and can be downloaded from `http://www.dynamictonality.com`.

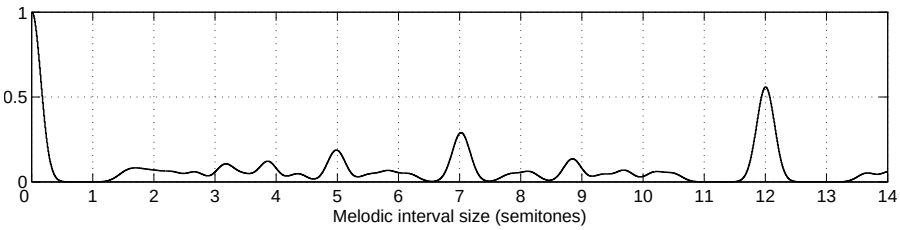


Figure 4.11: The spectral pitch similarity of harmonic complex tones with differing interval sizes. This chart is calculated with the parameter values for spectral similarity as optimized in Section 4.3.2.4 (smoothing of 10.3 cents, and roll-off of 0.42). The graph bears a clear resemblance to the sensory dissonance charts of, for example, Plomp and Levelt (1965) and Sethares (2005), with maxima of modelled affinity at simple frequency ratios like 2/1, 3/2, 4/3, and so forth.

val whose size is shown on the horizontal axis. Each tick corresponds to one 12-TET semitone, and a total range of just over one octave is covered.

Clearly, the intervals with the highest spectral pitch similarity (other than the unison) are the octave and the perfect fifth and perfect fourth. There is significant empirical evidence that the octave is universally recognized as an interval with extremely high affinity (Woolhouse (2009) cites numerous examples). The high spectral pitch similarity of perfect fifths and perfect fourths tallies nicely with historical evidence. For example, ancient Greek scales were typically based on conjunct and disjunct tetrachords. The two outer tones of a tetrachord span a perfect fourth (of frequency ratio 4/3) and, within this perfect fourth, lie two additional tones that could take on a wide variety of different tunings. The outer fourth was, however, always fixed. When a second tetrachord is placed a whole-tone above the top note of the first tetrachord (i.e., a perfect fifth above the bottom note), the entire octave is spanned to make a seven-tone scale. If the two tetrachords have identical internal structure, the resulting scale is rich in high spectral pitch similarity perfect fourths and perfect fifths. This technique of scale construction might, therefore, be seen as a heuristic for creating high-affinity scales. Indeed, the bounding fourths potentially provide perceptually secure start and end points for a melody that traverses the more challenging tones in between. For an in-depth examination of the history,

and mathematical, perceptual, and aesthetic properties of tetrachords, see Chalmers (1990), and for a discussion of the affinity (CDC-1) of the perfect fourth and fifth see Tenney (1988).

The diatonic and pentatonic scales, which are so ubiquitous to Western music, are the richest in terms of perfect fifths and fourths. This is because they are actually generated by a continuous chain of either of these intervals: there is no five-tone scale with more perfect fifths and fourths than the (anhemitonic) pentatonic, and no seven-tone scale with more perfect fifths and fourths than the diatonic. Such scales, therefore, maximize the number of the highest affinity (non-octave) intervals. Hypotheses that account for the development of such scales by the consonance of perfect fifths and fourths suffer from the uncomfortable historical fact that these scales were privileged well before harmony became commonplace in the West. Rephrasing this in terms of melodic affinity, rather than harmonic consonance, avoids this problem. It is also worth noting that Huron's (1994) analysis of the aggregate dyadic consonance of all possible scales selected from 12-TET indicates that the most familiar Western scales (the diatonic, ascending melodic minor and harmonic minor) have the highest aggregate consonance for 7-tone scales, the Japanese Ritsu and blues scale are highly ranked for 6-tone scales, and the anhemitonic pentatonic is the highest ranked 5-tone scale. This is relevant because affinity is correlated with consonance (see Figure 4.11), so similar results may hold if aggregated dyadic spectral pitch similarity were used instead of aggregated dyadic consonance. In this case, Huron's conclusions could be comfortably extended to scales developed for melodic music.

It would seem, therefore, that both historical and contemporary musical practice support the notion that spectral pitch similarity is a universal component of melodic affinity, and that this phenomenon has contributed to the types of scales that are, and have been, privileged in melodic music.

4.5 CONCLUSION

The experimental data give strong support for the hypotheses that the perceived affinity of successive tones is a monotonic function of both their spectral pitch similarity and their inherent harmonicity. The effect size of spectral pitch similarity is medium, the effect size of harmonicity is strong. The experimental intervention also suggests that spectral pitch similarity is modelling a nature process—one that does not require the learning of musical regularities.

In the next chapter, I build on this result and apply a related model to the relationships between tones and chords, and between chords and scales. In this way, I use spectral pitch similarity to model Krumhansl's probe tone data, and the stability and tonic-ness of chords in a variety of scales. I also present some hypotheses about how spectral pitch similarity may explain some additional aspects of tonality.

5

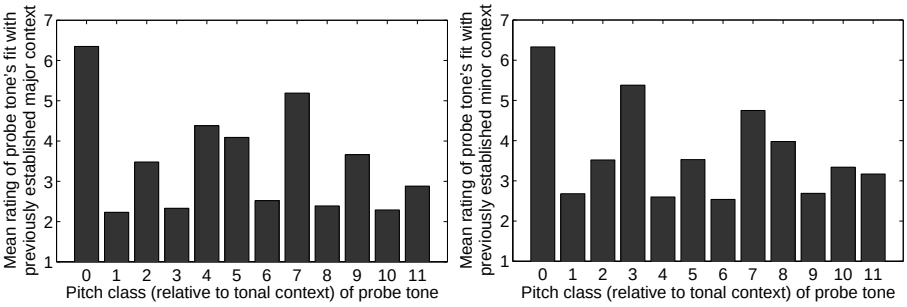
A MODEL OF THE PROBE TONE DATA AND SCALIC TONALITY

The Krumhansl and Kessler (1982) probe tone data are participants' ratings of the 'fits' of the twelve chromatic pitch classes to a previously established tonal context. They are widely considered to be one of the most important sets of experimentally obtained data about tonal perception. This is because, given a key, they can summarize the stability or tonic-ness of pitch classes, and they have a high correlation with the prevalences of pitch classes in Western tonal music. They have also proved effective at modelling perceived inter-key distance, and predicting the key of music as it plays. Furthermore, they are thought to be robust because similar results have been obtained in numerous related experiments.

Clearly, any proposed model of tonal perception should be able to effectively model this data. In light of the previously demonstrated success of the bottom-up spectral pitch similarity model at explaining the perceived affinity of microtonal pitches, it makes sense to test a related model for the probe tone data.

In this chapter, I test three spectral pitch similarity models of the probe tone data (they differ only in their parameterizations). I also compare my models against a number of previously suggested models, some of which are bottom-up, some of which are top-down. My own models have amongst the highest fit with the data and, being bottom-up, have wide explanatory scope.

Furthermore, I extend the model beyond the probe tone data and use it as a novel method to successfully predict the tonics of the diatonic scale and a variety of chromatic alterations of the diatonic scale (e.g., the harmonic minor and harmonic major scales). The model is also gen-



(a) Major context. (b) Minor context.

Figure 5.1: Krumhansl and Kessler's major and minor tonal hierarchies.

eralizable to any possible tuning, and I demonstrate this by predicting possible tonics for a selection of microtonal scales.

The probe tone data are commonly thought to represent *tonal hierarchies*—learned templates that assign fit levels to pitch classes once a specific major or minor key has been established (a deeper discussion is provided later in this introduction and in Sec. 5.2). However, because my model provides a bottom-up explanation for the probe tone data, it does not require learned templates. Such templates may exist, but my model is agnostic on this point. In reality, it is likely that a variety of long-term memory templates do play a supportive role and, in Section 5.2, I discuss the interplay between long-term memory (top-down) processes and bottom-up processes.

For the probe tone experiment, ten participants rated the degree of fit on a seven-point scale with 1 designated 'fits poorly' and 7 designated 'fits well' (Krumhansl and Kessler, 1982). These well-known results are illustrated in Figure 5.1.

The major or minor tonal context was established by playing one of four musical *elements*: just the tonic triad I; the cadence IV–V–I; the cadence II–V–I; the cadence VI–V–I. For example, to establish the key of C major, the chord progressions Cmaj, Fmaj–Gmaj–Cmaj, Dmin–Gmaj–Cmaj, and Amin–Gmaj–Cmaj were used; to establish the key of C minor, the chord progressions Cmin, Fmin–Gmaj–Cmin, Ddim–Gmaj–Cmin, and A♭maj–Gmaj–Cmin were used. A *cadence* is defined by Krumhansl and Kessler (1982, p. 352) as 'a strong key-defining se-

quence of chords that most frequently contains the V and I chords of the new key'; the above three cadences are amongst the most common in Western music. Each element, and its twelve probes, was listened to four times by each participant. For each context, the ratings of fit were highly correlated over its four different elements (mean correlations for the different elements were $r(10) = .90$ in major and $r(10) = .91$ in minor) so the ratings were averaged to produce the results shown in Figure 5.1. All listeners had a minimum of five years' formal instruction on an instrument or voice, but did not have extensive training in music theory.

All context elements and probes were played with *octave complex tones* (also known as OCTS or Shepard tones). Such tones contain partials that are separated only by octaves (i.e., they contain only 2^{n-1}th harmonics, where $n \in \mathbb{N}$), and the centrally pitched partials have a greater amplitude than the 'outer' partials; precise specifications are given in Krumhansl and Kessler (1982). Octave complex tones have a clear pitch chroma but an unclear pitch height; in other words, although they have an obvious pitch, it is not clear in which octave this pitch lies. The stated purpose of using OCTS was to 'minimize the effect of pitch height differences between the context and probe tones, which strongly affected the responses of the least musically oriented listeners in [an] earlier study' (Krumhansl, 1990, p. 26).

However, OCTS are unnatural acoustical events—no conventional musical instrument produces such spectra; they have to be artificially synthesized. Musical instruments typically produce *harmonic complex tones* (HCTS) in which most harmonics are present and such timbres contain a greater multiplicity of interval sizes between the harmonics (e.g., frequency ratios such as 3/2, 4/3, 5/3, and 5/4, in addition to the 2/1 octaves found in OCTS). Krumhansl and Kessler (1982, p. 341) describe the OCT timbre as 'an organlike sound, without any clearly defined lowest or highest component frequencies'.

The use of OCTS, rather than HCTS, may affect the resulting ratings of fit; that is, if HCTS had been used instead, it is possible the results may

have been—to some extent—different, even after taking account of pitch height effects (e.g., Parncutt (2011, p. 339) points out that the experimental data obtained by Budrys and Ambrazevicius (2008) indicates HCTs may reverse the fits of the minor third and perfect fifth—pitch classes 3 and 7—in the minor context). Despite this, the high correlation between the obtained ratings and the prevalences of scale degrees in samples of tonal music (e.g., $r(10) = .89$ for major and $r(10) = .86$ for minor (Krumhansl, 1990)) suggests that any such distortions are relatively small.

The probe tone data are thought to be important and robust. However, it should be borne in mind that only ten participants were involved, and they were all musically trained. The extent to which such results can generalize must, therefore, be open to some question. Parncutt (2011) cites a number of examples of similar experiments that have produced similar, but non-identical, results (Krumhansl and Shepard, 1979; Thompson, 1986; Steinke et al., 1993, 1998; Budrys and Ambrazevicius, 2008). Considering the probe tone data may be biased due to the small number of participants, and affected by the precise context-setting elements used, it is important that any model is judged not only on its goodness of fit, but also on the clarity and plausibility of its hypotheses, the extent to which it can explain the data, and its ability to make predictions beyond this single experiment.

The data are important because they can be generalized to predict aspects of music that were not explicitly tested in the experiment. Notably, the degree of fit can be used to model the stability or tonic-ness of the pitches and chords found in major-minor tonality—as originally suggested by Krumhansl (1990, pp. 16 & 19) and reiterated by Parncutt (2011, p. 333). Also, the data have been used to model perceived inter-key distances (Krumhansl and Kessler, 1982), and to predict the key—dynamically—of music as it plays (Krumhansl, 1990; Toiviainen and Krumhansl, 2003). However, there is no obvious way to use this data to account for certain important aspects of tonality: (a) Why is the primary major scale the diatonic, while the primary minor scale is

the non-diatonic harmonic minor scale?[1] (b) Why does the seventh degree (leading tone) of the major scale lose much of its activity when it is the fifth of the iii (mediant) chord? (c) Why are certain root progressions favoured over others (e.g., descending fifths are more common than ascending—particularly the cadential V–I)?

Krumhansl interprets the probe tone data as representing a *tonal hierarchy*. This can be thought of as a template that, given a tonal centre, induces degrees of perceived fit for all pitch classes (Krumhansl, 1990, pp. 16–21). For instance, when listening to music, pitch classes that are held in short-term memory are compared with all possible transpositions of the major and minor templates; the template whose profile is closest to the short-term memory pitch class distribution then induces appropriate levels of fit (stability, resolution versus tension, and rest versus activity) on the pitch classes. However, this is a hypothesis—the probe tone data may be, at least in part, a short-term response to the context elements (as argued by Butler (1989), Leman (2000), Parncutt (2011), and myself). For this reason, I avoid using the term *tonal hierarchies* to describe the data, preferring the neutral terms *probe tone data* and *fit profiles*. For clarity, when I am referring to a fixed and abstract tonal hierarchy, as envisaged by Krumhansl, I often use the term *template*.

In the Section 5.2, I challenge the conventional interpretation that the probe tone data necessarily represents a long-term memory *template*. In contrast, I suggest that—depending on the pitches held in short-term memory (i.e., the musical scale that is currently being used)—my bottom-up model can account *dynamically* for the fits and stabilities of pitches and chords without reference to a pre-formed template. As shown in a number of examples, when these pitches conform to familiar structures like the diatonic major scale, or the harmonic minor scale, the modelled fits are in accord with musical intuition and theory; when these pitches do not conform to these familiar structures (e.g., microtonal scales), alternative fit profiles are generated.

1 I use the term *diatonic* to refer exclusively to the scale with two steps sizes—L for large, and s for small—arranged in the pattern (L L s L L L s), or some rotation (mode) thereof (e.g., L s L L L s L, or s L L L s L L, etc.).

To provide the context for my model, in Section 5.1 I survey a variety of existing bottom-up and top-down models. In order to fairly compare the predictive power of the models (mine are nonlinear), I use cross-validation: my model comes in three variants and one of them is, to date, the best predictor of the data. But I am not concerned solely with their predictive power, I am particularly interested in their explanatory power: only my model and Parncutt's virtual pitch class model are both bottom-up and able to fit the data with exactitude. I am additionally interested in the generalizability of the model beyond the precise experiment it is modelling: both Parncutt's and mine are generalizable to any set of chromatically organized context and probes; my model is, additionally, generalizable to non-standard tunings—as I demonstrate in Section 5.3.2.

In Section 5.3.1, I additionally suggest some mechanisms that may help to answer the three previously-mentioned questions. These are that resolutions are strengthened when a worst-fitting pitch class moves to the root of a best-fitting triad, and that I also need to consider the fit of each pitch class within the chord it is part of. At the moment, however, these mechanisms are not instantiated in a formal mathematical model and, until they are, they should be thought of as preliminary findings or suggestions. I hope to formally embody these principles and test them against empirical data in future work (as discussed in Sec. 6.3).

5.1 THE MODELS

In this section, I provide a brief summary of my own and others' bottom-up and top-down models of the probe tone data. Most of these are also usefully summarized in Parncutt (2011), so I will keep my account brief. But I will also highlight a few areas where I take a different stance to Parncutt.

Before discussing each of the models in turn, Table 5.1 summarizes their relevant statistical properties with respect to the probe tone data. The correlations differ from those provided by Parncutt (2011) in two

Table 5.1: Correlations r, cross-validation correlations r_{CV} (both with 22 degrees of freedom), and cross-validation root mean squared errors of cross-validation RMSECV of the predictions made by a variety of models compared with Krumhansl and Kessler's (1982) probe tone data. The cross-validation statistics are the means of these statistics taken over twenty runs of 12-fold cross-validation (the cross-validation statistics are explained in Appendix E). The null model is an intercept-only model—i.e., all values are modelled by their mean. The remaining models are described in the main text. The models are ordered by their cross-validation statistics or, when these are missing, by their correlation.

	$r(22)$	$r_{\mathrm{CV}}(22)$	RMSECV	Type	Parameterization
Milne 12*c*	.97	.96	0.36	bottom-up	nonlinear
Lerdahl 88	.96	.95	0.38	top-down	linear
Parncutt 89	.96	.95	0.38	top-down	linear
Parncutt 94	.96	—	—	bottom-up	nonlinear
Parncutt 88/11*a*	.94	.93	0.45	bottom-up	linear
Milne 12*b*	.95	.92	0.48	bottom-up	nonlinear
Milne 12*a*	.94	.91	0.51	bottom-up	nonlinear
Parncutt 88/11*b*	.92	.90	0.54	bottom-up	linear
Smith 97	.89	.87	0.61	bottom-up	linear
Butler 89	.88	.84	0.65	top-down	linear
Krumhansl 90*b*	.87	.83	0.68	top-down	linear
Leman 00	$<.87$	—	—	bottom-up	nonlinear
Basic triad	.86	.82	0.70	—	linear
Krumhansl 90*a*	.65	.57	1.01	bottom-up	linear
Null	.00	$-.68$	1.27	—	linear

ways. Firstly, rather than providing one correlation value for the major context's probe tone data and one for the minor context's data, I give a single correlation value for both. I feel this is a more correct approach, because the same underlying process should apply to the major and minor contexts. Separately correlating them is equivalent to calculating the r-values of two linear regressions with different intercept and slope parameters. But there is no a priori reason to expect the two sets of parameters to be different, so I apply a single set of parameter values to both major and minor (in actuality, the correlations resulting from the two methods are almost identical—only one value differs by .01). Secondly, I additionally provide cross-validation statistics (r_{CV} and RMSECV, for which higher values of the former and lower values of the latter indicate a better model)—the reason for this is to allow my nonlinear models to be fairly compared with the mostly linear mod-

els that have been proposed so far. As described in Section 2.2, utilizing un-cross-validated statistics would not be appropriate, because the additional flexibility of a model with additional nonlinear parameters may allow it to fit the noise rather than the process underlying the data, thereby giving it an unwarranted advantage. Cross-validation statistics provide a way for models with differing levels of flexibility (complexity) to be fairly compared, and ensure they are not overfitting the data.[2] I give a more technical explanation of the cross-validation statistics in Appendix E.

It is worth pointing out that the modelled data do not need to replicate much of the experimental data's fine structure in order to achieve what appears to be a reasonably good correlation value. For example, let us define a *basic triad model* as one that gives the tonic chords' pitches a value of 1, and all other pitch classes a value of 0; the resulting statistics are surprisingly impressive looking: $r(22) = .86$, $r_{\mathrm{CV}}(22) = .82$, and $\mathrm{RMSECV} = 0.70$. I suggest that any model with similarly valued statistics is probably struggling to describe much of the fine structure of the data; I have placed this basic triad model into Table 5.1 to serve as a benchmark.

5.1.1 *Krumhansl 90b: corpus prevalence model.*

Krumhansl (1990) suggested a predictively effective model for the probe tone data ($r_{\mathrm{CV}}(22) = .83$). The hypothesis is that the data are correlated with the distribution (prevalences) of scale degrees in existing music. This is a purely top-down model of music perception, in that the perceived fits of the probe tones are hypothesized to be down to nothing more than learning: if we frequently hear the fourth scale degree, we will tend to feel that scale degree has a good fit; if we rarely hear altered scale degree $\flat\hat{2}/\sharp\hat{1}$, we will tend to feel that scale degree has a poor fit. (To reiterate from Section 2.3, a *top-down* model is one that utilizes, as variables, the prevalences of musical patterns or the preva-

2 Analytic approaches such as AIC and BIC are not appropriate here because the degrees of freedom for my model cannot be estimated.

lences of their associations with non-musical events; in other words, top-down models require a statistical analysis of a musical corpus.)

This model provides a straightforward explanation for our perception of scale degree fit, but the scope of this explanation is limited. It cannot explain why the probe tone data/scale degree prevalences take the specific profile they do. Indeed, an implicit assumption of this model is that this profile is down to nothing more than chance—for some unknown reason, composers favoured certain scale degrees and hence listeners came to feel these scale degrees fitted better. Composers (who are also listeners) continued to write music that utilized these learned patterns of fit (because such music made sense to them and their listeners), and so listeners (some of whom are composers) continued to have their learning of these patterns reinforced. And so forth, in a circular pattern of causal effects: music perception is the way it is because music is the way it is, and music is the way it is because music perception is the way it is, ad infinitum. Presumably, this theory predicts that on a 'parallel Earth'—identical in all respects to ours except for random fluctuations—a completely different profile of pitch class fits might have developed.

Of course, this may be true. But it is quite plausible that there are innate perceptual or cognitive principles that might contribute to making one, or a small number, of actual fit profiles possible or more likely. Any theory that can provide a bottom-up explanation for why the probe tone data have the specific forms we observe, has greater explanatory power than a theory that is purely top-down. (To reiterate, a *bottom-up* model is one that does not utilize, as variables, the prevalences of musical patterns or the prevalences of their associations with non-musical events; in other words, bottom-up models do not require a statistical analysis of a musical corpus.)

In the subsequent model descriptions, I shall point out whether they are bottom-up or top-down: hence I can discuss their explanatory power as well as their as their predictive power.

5.1.2 *Lerdahl 88: pitch space model.*

Lerdahl's *basic pitch space* is a model for tonal distances (Lerdahl, 1988). It has five levels: (1) tonic, (2) tonic and fifth, (3) major tonic triad, (4) diatonic major scale, (5) chromatic scale. He points out that the five levels in this basic pitch space correlate well with the major context's probe tone data. He does not, however, suggest a formal model for the minor context. To address this, it is necessary to create a conceptually related 'minor pitch space' for the minor context. Lerdahl's model (and its extension to the minor context) is predictively very effective ($r_{\mathrm{CV}}(22) = .95$). However, it is deficient in terms of explanatory power because important aspects of the basic pitch space itself are derived from (or require) top-down explanations.

Lerdahl provides a bottom-up explanation for the first three levels, which is that the height of a level should correlate with 'the degree of sensory consonance of adjacent intervals' within it (Lerdahl, 2001, p. 272) (he defines sensory consonance psychoacoustically as a function of both roughness and clarity of the root (2001, p. 321)). The perfect fifth in the second level is the most consonant interval, and the major triad on the third level is the most consonant triad (although the minor triad is similarly consonant and seems a reasonable alternative). The fourth level—which is critical for producing high correlations with the data—is the diatonic major scale. Although Lerdahl gives a number of bottom-up explanations for privileging the diatonic scale,[3] he gives only a top-down explanation for choosing its Ionian (i.e., major) mode, rather than the Mixolydian or Lydian—he privileges the former due to its prevalence (2001, p. 41). The predictive power of the basic pitch space, therefore, relies on a long-term memory explanation, so I class this model as top-down.

To extend Lerdahl's model to account for the minor context, Parncutt created a 'minor pitch space'. This builds up the levels in the same way, but has a minor triad (rather than a major triad) on the third level,

3 Balzano's principles of uniqueness, coherence, and simplicity, and Clough and Douthett's maximal evenness (Lerdahl, 2001, pp. 50–51 & p. 269).

and has the harmonic minor scale (rather than the diatonic major scale) on the fourth level. The resulting major (basic) and harmonic minor pitch spaces are highly correlated with their respective probe tone data ($r_{\mathrm{CV}}(22) = .94$) (this version of Lerdahl's model is not included in Tab. 5.1).

However, in one respect, this harmonic minor pitch space is not in keeping with Lerdahl's conceptualization of the basic pitch space because it uses a non-diatonic scale (the harmonic minor), which does not have the property of coherence, for the fourth level. It is arguably more in keeping with Lerdahl's theory to use the coherent Aeolian mode (natural minor scale) for the fourth level. This Aeolian pitch space actually has a higher correlation with the minor context's data than Parncutt's harmonic minor version, and it is this Aeolian version of Lerdahl's model that I include in Table 5.1. This model is predictively extremely effective ($r_{\mathrm{CV}}(22) = .95$). However, as an essentially top-down model, it has limited explanatory power.

5.1.3 *Butler 89: aggregate context pitch multiplicity model.*

Butler (1989) presents his model as utilizing nothing more than short-term memory, in which case, it is an explanatory bottom-up model. However, as we shall see, it is actually more likely that this is a top-down model of a possible long-term memory process.

He models the probe tone ratings simply by the number of times their pitches occur in each context's elements (i.e., the chord progressions I, IV–V–I, II–V–I, and VI–V–I). These four elements were aggregated into a chord collection containing IV, II, VI, three Vs, and four Is. The model counts the number of occurrences of each scale degree in this collection: there are six $\widehat{1}$s (in the four Is, the IV, and the VI); there are zero $\sharp\widehat{1}/\flat\widehat{2}$s; there are four $\widehat{2}$s (in the three Vs and the II); and so on. The resulting counts for the major and minor contexts' elements fit the data well ($r_{\mathrm{CV}}(22) = .84$). As a short-term memory model, it is bottom up and provides a meaningful explanation for why,

given an immediate context element, certain pitches (probes) fit better than others: currently heard pitches that are also salient in short-term memory are perceived to fit better than pitches that are not also salient in short-term memory—we are 'comfortable' with, or 'less surprised' by, repetition. It also implies that there is not necessarily a stable tonal hierarchy that serves as a fixed template against which currently heard pitches are compared.

However, it is questionable whether this model can be considered to be a short-term memory model. As Krumhansl (1990, p.62) points out, the different context elements were presented to listeners in separate blocks, not intermixed within the same block and, for this reason, it is implausible that short-term memory—which typically completely decays within 20 seconds (Peterson and Peterson, 1959)—could be responsible for aggregating the four elements (this point is also amplified by Woolhouse and Cross (2010)). If Butler's model is applied to each context element separately and then averaged over them, the fit with the probe tone data is substantially poorer (averaged $r_{\mathrm{CV}}(22) = .74$ and averaged $\mathrm{RMSECV} = 0.79$). So, when corrected to more accurately reflect short-term memory processes, the model becomes predictively weak.[4] Furthermore, Krumhansl and Kessler (1982, p. 343) found the ratings produced by the differing context elements to be 'very similar', whereas the modelled data produced by the differing context elements is not.

As pointed out by Parncutt (2011, p. 341), a mechanism that could account for the aggregation of the four context elements being correlated with the data would be that the aggregated chord context is a good summary of the prevalences of chords in Western music. However, this transforms the model into a purely top-down model, where the fit of probe tones is solely down to their prevalence. In other words, viewed from this perspective, Butler's model is the same as Krumhansl's preva-

4 The only practicable way to perform the cross-validations was to allow for the parameters, within each training fold, to vary across the different context elements. There is, however, no a priori reason why they should be different over different context elements. If they had have been kept the same, the resulting statistics would have been even lower.

lence model; the difference being that Krumhansl statistically analyses a corpus, while Butler statistically analyses a set of common cadences—and both have similar scale degree prevalences. For this reason, I class this model as top-down.

5.1.4 *Parncutt 89: aggregated context pitch class salience model.*

Parncutt (1989) adapted Butler's model in two ways. Firstly, he used a different aggregation of the contexts' elements: IV, II, VI, three Vs, and six Is. The difference is that the tonic triad element is counted six rather than four times, this is because Parncutt counts the tonic triad three times for the context element that comprises only the I chord. Despite Krumhansl's criticism (1990, p.62) that this does not reproduce the stimuli used in the experiment, it is actually quite reasonable because the ratings produced by the four context elements were averaged to produce the final sets of probe tone data (so, counting the I element three times, gives it equivalent weight to each of the other three elements (Parncutt, 1989, p. 159)). Secondly, he included not just the notated pitches in the context elements, but also their pitch class (or chroma) salience profiles. The precise mechanism by which the pitch class saliences are generated for a harmonic complex tone is detailed in Parncutt (1989, Sec. 4.4.2). In summary, the salience of any given pitch class is calculated from a combination of the weights of harmonics and subharmonics with corresponding pitch classes—these subharmonics and harmonics extending from each notated pitch. The subharmonics are, overall, weighted significantly higher than the harmonic pitches, so this is primarily a virtual (subharmonic) pitch model.

The model produces one of the best fits to the data ($r_{\mathrm{CV}}(22) = .95$), but it suffers from the same problems as Butler's: it cannot really be interpreted as a model of short-term memory processes; rather, it is a model of a possible long-term memory process, where the aggregated cadences serve as proxies for prevalent chords in Western music. So the model has limited explanatory scope—although it may explain the

data given the prevalence of a small set of chords, it does not explain why those chords, in particular, are prevalent.

5.1.5 *Leman 00: short-term memory model.*

Leman (2000) utilizes a short-term memory model whose inputs are derived from a model of the auditory system. The latter comprises forty bandpass filters, half-wave rectification and simulations of neural firings induced by the filters, and periodicity detection (autocorrelation) applied to those firings. Autocorrelation automatically detects frequencies that are subharmonics of the input frequencies. In this respect it is, therefore, similar to the above Parncutt 89 model. The resulting signals, produced in response to the context element, are stored in a short-term (echoic) memory model which decays over time and, at the time at which the probe is presented, this represents the 'global image' of the context element. The length of the decay (the half-life of the signal) is a free parameter. This global image is correlated with a 'local image' produced by each of the 12 probe tones (for each of the four context elements in both major and minor). The twelve correlation values (for the twelve probes) are averaged over the four major and four minor context elements (in the same way as Krumhansl's data), and these are used to model the probe tone data.

The model produces correlations towards the lower end of those discussed here ($r(10) = .85$ for major and $r(10) = .83$ for minor). However, Leman chooses a decay parameter of 1.5 seconds, when his Table 3 shows that the maximum decay value tested (5 seconds) would have fit the probe tone data better (he chooses the lower time value because fitting the probe tone data is not his only criterion). With the 5 second decay time, the correlations improve, but only slightly ($r(10) = .87$ for major and $r(10) = .84$).

Because of the nonlinear decay time parameter, and without easy access to the original model, I have not calculated its cross-validation correlations, and root mean squared errors of cross-validation. However,

since the $r(22)$ statistics will be lower than .87 (which is the highest $r(10)$ statistic gained by the 5 second decay time model of the major context's data), it is safe to conclude that, in terms of prediction, this is one of the worst performing models and probably no better than the 'basic triad' benchmark model.

5.1.6 *Krumhansl 90a: consonance model.*

Krumhansl's (1990) other model is bottom-up and attempts to provide a more substantive explanation than the prevalence model. It also predicts rather poorly ($r_{\mathrm{CV}}(22) = .57$). This model hypothesizes that the probe tone fits are due to the consonance of the corresponding pitch class and the tonic pitch class (the first scale degree). Clearly, this model will struggle to obtain high correlations with the empirical data because it produces identical predictions for the major and minor contexts (they both have the same tonic pitch class).

Krumhansl uses consonance values that are the averages of a variety of bottom-up models of consonance (Malmberg (1918), Helmholtz (1877), Hutchinson and Knopoff (1978), and Kameoka and Kuriyagawa (1969)), and one set of empirically derived consonance ratings (Malmberg (1918)). This means the model, as a whole, is essentially bottom-up and has wide explanatory scope—it provides an explanation for the probe tone ratings based on innate perceptual processes. However, it is also worth noting that—as Krumhansl points out (1990, p. 55)—there is something of a mismatch between the model's explanation and the experimental procedure used to get the empirical data: the probe tones were played after the context-setting chords, not simultaneously, so harmonic consonance/dissonance does not play a direct role in the experimental stimuli. For this model to make sense, it must be additionally assumed that the listeners were mentally simulating harmonic intervals comprising the tonic and the probe, and then determining their consonance/dissonance values either directly or from long-term mem-

ory. This is plausible, given the musical experience of the participants, but it is an indirect explanation.

5.1.7 *Smith 97: cumulative consonance model.*

Like Krumhansl, Smith (1997) also uses consonance—but in a different way—to explain the data from the bottom up. He takes a tonic pitch and finds a second pitch with the greatest consonance. To these two pitches, he then finds the third pitch that makes the most consonant three-tone chord (in all cases, consonance is calculated as the *aggregate dyadic consonance*, which is the sum of the consonances of all interval classes in the chord (Huron, 1994)). To this three tone chord, he finds the pitch of the fourth tone that creates the most consonant four-tone chord. And so forth, until all 12 pitch classes are utilized.

If the first pitch is C, the second pitch is G, and the third pitch is either E or E♭ (the major and minor triads have equal aggregate consonance because they contain the same three interval classes, 3, 4, and 5). Because there are two possible three-tone chords, the resulting cumulatively constructed scales bifurcate at this juncture. For the major triad C–E–G, the fourth tone is A; for the minor triad C–E♭–G, the fourth pitch is B♭. Continuing this process, leads to the following two sequences of pitch classes: C–G–E–A–D–F/B–A♭–G♭/B♭–D♭/E♭, and C–G–E♭–B♭–F–D/A♭–B–D♭/A–E/G♭ (where X/Y denotes that X and Y have the same ranking). When each pitch class is assigned a value according to its ranking (e.g., in the first sequence, $\mathrm{C} = 1$, $\mathrm{G} = 2$, $\mathrm{E} = 3$, $\mathrm{A} = 4$, $\mathrm{D} = 5$, $\mathrm{F} = 6.5$, $\mathrm{B} = 6.5$, $\mathrm{A}\flat = 8$, etc.), they provide a predictively effective model of their respective major and minor probe tone ratings ($r_{\mathrm{CV}}(22) = .87$).

This model has reasonable predictive power (though its predictive performance is towards the lower end of the models discussed here) and, like Krumhansl's 90*a* consonance model, has potential for good explanatory power if the consonance values it uses are derived from a psychoacoustic or other bottom-up model. Smith actually uses in-

terval class consonance values derived by Huron (1994) from empirical data collected by Kameoka and Kuriyagawa (1969), Hutchinson and Knopoff (1978), and Malmberg (1918), not from modelled data. Using empirical data means that the consonance values are likely to be correct and do not have to rely upon possibly inaccurate models (Huron, 1994). However, this weakens the explanatory scope of Smith's model—ideally, a bottom-up consonance model would be substituted at some stage. Like Krumhansl's consonance model, this model also suffers from the indirect relationship between harmonic consonance (the model's variables) and melodic fit (what the experiment actually measures).

5.1.8 *Parncutt 88/11 & 94: virtual pitch class models.*

Parncutt provides the first predictively effective bottom-up model I have described so far ($r_{\mathrm{CV}}(22) = .93$). It builds on Parncutt's (1988) model of virtual pitch classes, and the concept of 'tonic as triad', which is explored in Parncutt (2011). (The model described here was first presented in 2011, though aspects of it date back to 1988.) This concept treats the tonic as a triad—a major or minor chord built upon the tonic pitch class—and it can be seen as a break from a more traditional concept of 'tonic as pitch class'.[5] For example, the tonic of the key C major is not the pitch class C, but the triad Cmaj; the tonic of the key B♭ minor is not the pitch class B♭, but the triad B♭min.

The tonic as triad concept implies that the context-setting elements—whose purpose is to induce a strongly defined key and all of which end in the tonic triad—can be effectively represented by the tonic triad. For instance, the cadence Fmaj–Gmaj–Cmaj is used to establish the chord Cmaj as a strong and stable tonic chord, so it is unsurprising if our attention is more clearly focused on the Cmaj chord than on the preceding chords. Indeed, even if the elements were, for example, Fmaj–Gmaj, or only G7, even though the Cmaj is not actually

5 An early description of the tonic as triad concept is given in Wilding-White (1961).

played it is still easy to imagine it as the most expected (and best fitting) continuation. The tonic triad, therefore, effectively summarizes our response to the context-setting elements used in the experiment; importantly, it also effectively summarizes our response to tonal context-setting devices in general. Henceforward, this tonic triad is simply referred to as the *tonic triad context*.

The probe tone ratings are modelled from the weights of the *virtual pitches* that are internally generated in response to the notated pitches in the tonic triad context. (By *internally generated*, I mean that virtual pitches are produced by some aspect of the auditory or cognitive system—they are not physically present in the stimulus prior to entering the ear.) Virtual pitches are typically modelled to occur at subharmonics below the notated pitch (the first N subharmonics of a notated pitch with frequency f occur at frequencies $f, f/2, f/3, \ldots, f/N$). There is well-established evidence that virtual pitches are generated from physical pitches—for example, if the fundamental is physically removed from a harmonic complex tone, its pitch still heard as corresponding to that missing fundamental, and combination tones produced by multiple sine waves are clearly audible phenomena. However, the extent to which HCTs (or OCTs) produce salient virtual pitches at pitch classes different to that of their fundamental is less obviously demonstrable.

In Parncutt's model, the pitch of each subharmonic is modelled in a categorical fashion; that is, it is categorized by the pitch class it is closest to. For example, the seventh subharmonic below C4 corresponds to a pitch 31 cents above D1, but is modelled by the pitch class (category) D. The model, therefore, hypothesizes that pitch discrepancies of the order of a third of a semitone have no impact on whether that pitch is mentally categorized as a specific chromatic pitch class.[6] For any given notated pitch, its virtual pitch classes are weighted: the virtual pitch

6 Parncutt (1988, p. 70) argues such pitch differences can be ignored because the seventh harmonic of an HCT can be mistuned by approximately half a semitone before it sticks out; conversely, I would suggest that when musicians' pitches go off by more than about 20 cents, the notes are generally perceived as out-of-tune, and do not comfortably belong to their intended (or any other) chromatic pitch class category.

class corresponding to the notated pitch class itself has weight 10; the virtual pitch class seven semitones (a perfect fifth) below has weight 5; the virtual pitch class four semitones (a major third) below has weight 3; the virtual pitch class ten semitones (a minor seventh) below has weight 2; the virtual pitch class two semitones (a major second) below has weight 1. These weights are justified on the grounds that they are numerically simple and are approximately proportional to the values achieved by taking a subharmonic series with amplitudes of $i^{-0.55}$, where i is the number of the subharmonic (a typical loudness spectrum for the harmonics produced by musical instruments), and summing the amplitudes for all subharmonics with the same pitch class (Parncutt, 1988, p. 74).

These virtual pitch classes, and their weights, are applied to the three notated pitches in the major or minor tonic triad; when virtual pitch classes from different notated pitches are the same, their weights are summed to model the overall virtual pitch class weights produced by a tonic triad. For example, in the chord Cmaj, the notated pitch C contributes a virtual pitch class C of weight 10, the notated pitch G contributes a virtual pitch class C of weight 5, the notated pitch E contributes a virtual pitch class C of weight 3; the three are combined to give a virtual pitch class C with a total weight of 18. The two sets of virtual pitch class weights for a major and minor triad closely fit their respective probe tone data, and do so with a plausible bottom-up (psychoacoustic) model.

A natural explanation provided by this model would appear to be that the greater the commonality of the pitches evoked by the tonic triad (which represents the context) and those evoked by the probe, the greater the perceived fit. However, in Parncutt's model (which is designated Parncutt 88/11a in Table 5.1), the probe tone itself is modelled with a single pitch, rather than as a collection of virtual pitch classes. It is not clear why the tonic triad context should evoke virtual pitches, but the probe does not; the probe's missing virtual pitches seems like a conceptual inconsistency in this model. If the probe tone is given vir-

tual pitch classes—in the same way as the tonic triad context—the resulting predictions are still good, but slightly less accurate ($r_{\mathrm{CV}}(22) = .90$). This is shown as Parncutt 88/11*b* in Table 5.1.

It is worth pointing out that in an earlier model, Parncutt (1994) utilized a similar virtual pitch class model that included all of the chords played in each context-setting element, but adjusted their weights to account for short-term memory decay (similar to that described for Leman 00). The memory half-life was a nonlinear parameter optimized to 0.25 seconds; this means the model incorporates the virtual pitch classes of the final tonic, and—to a lesser degree—the virtual pitch classes of the preceding chords. Interestingly, this model performs well for each context-setting element (unlike Butler 89 and Parncutt 89) and, when averaged across the elements, has a slightly better correlation than the Parncutt 88/11*a* model—as shown in Table 5.1, where it is designated Parncutt 94. I was unable to calculate the cross-validation statistics because I do not have access to the original model, but they are unlikely to be significantly better than Parncutt 88/11*a*. These results suggest that utilizing all the chords in a given context-setting element works well for predicting the response specific to that element, but using just the tonic triad is sufficient for capturing the effects of harmonic tonality more generally; that is, averaged over a broader range of chord progressions.

The bottom-up explanation provided by these models (pitch class commonality of the context and probe) can be generalized beyond the experimental setup, in that we might imagine that, when listening to a piece of music, we mentally test the virtual pitch class commonality of a variety of potential tonic pitches or tonic triads to the elements held in short-term memory. The best-fitting pitch functions as the local melodic tonic, the best-fitting triad functions as the local tonic triad. This concept was utilized in Parncutt and Prem (2008) and Parncutt (2011) as a possible explanation for which diatonic modes were favoured in medieval music; though, in these examples, the abovementioned inconsistency applies in that virtual pitch classes are given

only to one half of the comparison—the (context) diatonic scale's notated pitch classes have virtual pitches, the (probe) candidate tonics' notated pitch classes do not. In Section 5.3, I explore this notion in greater depth and provide the modelled fits generated by a variety of scales using an approach that is consistent with respect to the context and the probe.

5.1.9 *Milne 12: spectral pitch class similarity models.*

For my models, I use Parncutt's central insight of the tonic as triad, but I use a different measure of the similarity between the probe tones and this tonic—I use *spectral pitch class similarity* rather than virtual pitch class commonality. This is the same as the method used in Chapter 4, except that I embed the spectral pitches as pitch classes. Pitch classes are used rather than pitches because I am interested in modelling the perceived distances of pitch classes. For example, I want to know the distance between the notes C and F, regardless of whether the former is above or below the latter.

Like the model in Chapter 4, this one is also parameterized by a roll-off ρ that models the relative importance of higher partials (the amplitude of the ith harmonic is given by $i^{-\rho}$), and a pitch difference limen of σ cents that models the inaccuracy of pitch perception (for a full mathematical definition of the model, see Appendix F; the MATLAB routines can be downloaded from `http://www.dynamictonality.com/probe_tone_files/`).

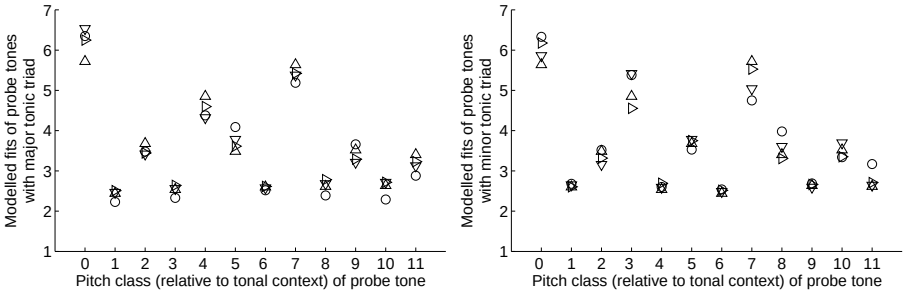
In each of my three models I allow for different weightings of the tonic triads' pitches. In model a, I give all their pitches the same weights—that is, the saliences of the partials in its three pitch classes, as previously determined by ρ, are multiplied by 1 and so left unchanged. In model b, two weightings are available—the tonic triads' roots have unity weight, while the remaining pitch classes have a weight of ω, which takes a value between 0 and 1; for example, if the tonic triads are Cmaj and Cmin, the saliences of the partials of the pitch class C are left

unchanged, while the saliences of the partials of all the remaining pitch classes are multiplied by ω. In model c, there are still two weightings, but this time the unity weight is applied to the roots of the major and minor tonics and also the third of the minor tonic, while the weighting of ω is applied to the remaining pitch classes; for example, if the tonics are Cmaj and Cmin, the weights of the partials of the pitch classes C and E♭ are unchanged, while the weights of the remaining pitch classes are multiplied by ω.

Model a is a pure tonic-as-triad model (all its three pitch classes are equally weighted), but the separate weightings in b and c allow these models to be situated in continua between tonic-as-triad and tonic-as-pitch models. This is useful because it is plausible that, of the tonic triad's pitches, the tonic pitch is the most salient and tonic-like. Model c treats the third of the minor triad as an additional root and as a frequent substitute tonic. This is in accord with musical practice, which often treats the third of the minor chord as a stable root (minor chords in first inversion are not treated as dissonances) and, in minor keys, modulations to the relative relative major are very common (the tonic of the relative major is the third of the minor tonic's triad).

Therefore, in addition to the intercept and slope parameters (which are part of every model discussed so far due to the process of obtaining correlation values),[7] model a has two additional nonlinear parameters (ρ and σ), while models b and c have three additional nonlinear parameters (ρ, σ, and ω). This nonlinearity means the parameter values cannot be optimized analytically, so I used MATLAB's `fmincon` routine to optimize them iteratively. I optimized each model so as to minimize the sum of squared errors between its predictions and the probe tone data—this is the same for all the models discussed in this chapter, because obtaining correlation values automatically chooses intercept and slope values that minimize the sum of squared errors.

7 The correlation coefficient between a model's data and the empirical data is given by $\sqrt{(\widehat{\mathbf{y}}-\mathbf{y})'(\widehat{\mathbf{y}}-\mathbf{y})/(\mathbf{y}-\overline{\mathbf{y}})'(\mathbf{y}-\overline{\mathbf{y}})}$, where $'$ is the transpose operator which turns a column vector into a row vector, $\mathbf{y}$ is a column vector of the empirical data, $\overline{\mathbf{y}}$ is a column vector all of whose entries are the mean of the empirical data and, critically, $\widehat{\mathbf{y}}$ is a column vector of the model's predictions *after* having been fitted by simple linear regression.



(a) Major context. (b) Minor context.

Figure 5.2: The circles ○ show the probe tone data, the upwards pointing triangles △ show the data as modelled by model a, the rightwards pointing triangles ▷ show the data as modelled by model b, the downwards pointing triangles ▽ show the data as modelled by model c.

The optimized parameter values all seem plausible: for model a, $\widehat{\rho} = 0.52$ and $\widehat{\sigma} = 5.71$; for model b, $\widehat{\rho} = 0.77$, $\widehat{\sigma} = 6.99$, and $\widehat{\omega} = 0.63$; for model c, $\widehat{\rho} = 0.67$, $\widehat{\sigma} = 5.95$, and $\widehat{\omega} = 0.50$.[8] The values of ρ are all similar to the loudnesses of partials produced by stringed instruments (a sawtooth wave, which is often used to synthesize string and brass instruments, has a pressure roll-off equivalent to a ρ of 1 and, using Steven's law, this approximates to a loudness roll-off equivalent to $\rho = 0.6$). Under experimental conditions, the frequency difference limen (just noticeable difference) corresponds to approximately 3 cents, which would be modelled by a smoothing width of 3 cents (as explained in App. A). In a musical experiment like the one being modelled, we would expect the smoothing to be somewhat wider, and the value of around 6 cents seems plausible. Both $\widehat{\rho}$ and $\widehat{\sigma}$ are similar to those obtained in Chapter 4 (which were 0.42 and 10.28, respectively). Their similarity is demonstrated by the fact that utilizing them instead has only a small negative impact on the resulting fit (reducing the correlation values by approximately 0.003). This also indicates that the model is robust over such changes to these parameters.

The optimized spectral pitch class similarity models are predictively effective—for models a, b, and c, respectively, the cross-validation

8 With iterative optimization, there is always a danger that a local rather than global minimum of sum of squared errors is found; I tried a number of different start values for the parameters, and the optimization routine always converged to the same parameter values so I am confident they do represent the global optimum.

statistics are $r_{\mathrm{CV}}(22) = .91$, $r_{\mathrm{CV}}(22) = .92$, and $r_{\mathrm{CV}}(22) = .96$. The predictions made by the three models are shown in Figure 5.2. They also have great explanatory power—like Parncutt's virtual pitch class model, we are using psychoacoustic principles to explain the specific shape taken by the probe tone data.

However, there is one aspect of these models that does not bear a direct relationship with the experimental procedure. In the experiment, the stimuli were all OCTS, not HCTS. In my models, I use HCTS (if OCTS are used as variables, the models perform very poorly). (This is also the case in Krumhansl's and Smith's consonance models, because their consonance values are all derived from HCTS.) There are at least four possible explanations that can bridge the gap between the model's use of HCTS and the experiment's use of OCTS. Firstly, nonlinearities in the auditory system may add harmonics to the OCTS (e.g., a combination tone of any two adjacent OCT partials with frequencies f and $2f$, has a frequency at $3f$—a third harmonic). Secondly, when listeners were making their judgements of fit, the representations of the tonic triad context and probe they retrieved from short-term memory may have been 'contaminated' by long-term representations of HCTS with the same pitch (HCTS being much more familiar). Thirdly, listeners may have recalled the levels of fit, stored in long-term memory, of equivalently sized HCT intervals. Fourthly, listeners' judgements of the fit of the probe and the tonic triad context are due to musical prevalence, but these musical prevalences are themselves a function of the psychoacoustic process modelled here: specifically, composers usually work with HCTS (not OCTS) and build up a set of tonal prevalences based upon their desire to follow their innate and universal perceptual processes (and 'consumers' support music that accords with their similar innate processes). Although the last explanatory bridge includes a top-down component, it is still fully explanatory because the bottom-up (spectral pitch class similarity) component explains why the fit profiles take the specific form they do.

5.2 A TEMPLATE-INDEPENDENT THEORY OF TONALITY

Krumhansl interprets the probe tone profiles as representing learned hierarchies—templates—of the differing fits of all twelve pitch classes to a previously identified tonal centre; when listening to a piece of music, we compare the prevalences of pitches held in short-term memory with the profiles of all twelve transpositions of the major and minor templates held in long-term memory (this is the process modelled by Toiviainen and Krumhansl (2003) to predict the perceived key of a piece of music—dynamically—as it progresses). The hierarchy whose profile most closely matches the short-term pitch prevalences then induces its levels of fit and stability on all twelve pitch classes. At any given moment, therefore, our perceptions of pitch class fit and stability are completely determined by one of these fixed templates.

However, if there is a bottom-up explanation for the probe tone data, templates are no longer necessary to explain the origin of perceived fit and stability. Instead, we can hypothesize that, given a collection of pitches in short-term memory, we are able to mentally 'calculate' or 'feel' which triad is the best fitting, and that this triad is then perceived as the most stable and tonic-like. This is a somewhat more dynamic conception of tonal perception, in that the tonic and its relationship to all twelve pitch classes is not assumed to be fixed: rather, the tonic-ness of any chord is assessed constantly on the basis of the pitch classes that have been recently played. Indeed, Butler (1989), Leman (2000), and Parncutt (2011) have all argued that processes based purely upon short-term memory play a role in our assessment of fit. For instance, Parncutt (2011, pp. 338–339) points out that, although experimental observations of probe tone data are fairly consistent over differing context-setting elements, they are not identical.

Having said that, it would be implausible to completely dismiss the impact of long-term memory. It is likely that certain scales (e.g., the diatonic and harmonic minor) are so commonly used that we learn where the best fitting chords are without having to mentally assess it

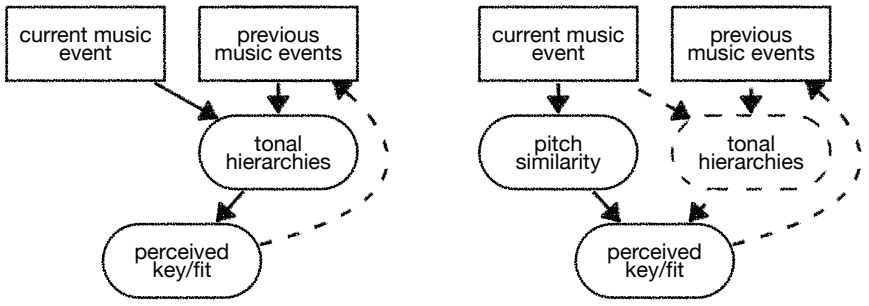
every time, and that composers' usages of chords follow these innate tendencies, which further amplifies their patterns of fit. It is also likely that we become familiar with specific sequences (ordered sets) of pitch classes and chords, which are able to most clearly exemplify the patterns of fit implied by a scale. For example, as I discuss in the examples below, movements from chords containing pitch classes with low fit to those with high fit may provide particularly effective resolutions, which strongly define a tonic. This means that any long-term memory templates that induce stability, resolution, activity, and so forth, may be quite diverse in form, consisting of a variety of pitch and chord-based fragments rather than just two overarching hierarchies.

Even allowing for such long-term memory processes in bottom-up theories, a key difference still remains: top-down (template) theories cannot explain the origin of the templates themselves, whereas bottom-up (template-independent) theories have a causal explanation for why certain chords and pitches function the way they do, and for the specific form taken by any templates that may arise through the prevalence of certain scales, melodies, and chord progressions that exploit innate perceptual processes for aesthetic ends. In Figure 5.3, I replicate Figure 2.4, but substitute the processes and variables relevant to this experiment. This allows the causal flows implied by these different types of models to be clearly illustrated.

A bottom-up model that rejects the exclusivity of long-term memory templates has an important implication: we can predict tonics on the basis of specific contexts such as scales, and it enables us to talk of a *scalic tonality* whereby any unique collection of pitch classes has unique tonal implications—even if not previously heard. I explore this in the following section.

5.3 SCALIC TONALITY

In this section, I describe a method to predict the levels of fit and stability for pitches and chords in any given scale. The method can be



(a) Top-down (template) models. This type includes Lerdahl 88, Butler 89, Krumhansl 90b. The differences between these models is their method for estimating the statistical distribution of previous musical events. Composers may or may not be hypothesized to provide a causal flow from music perception to the repertoire (hence the dashed arrow).

(b) Bottom-up (template-independent) models. This type includes Parncutt 94, Leman 00, Parncutt 88/11, Milne 12. Pitch similarity refers to the perceived similarity of the context and the probes. The different models use different methods for determining their pitch contents and similarities. Tonal hierarchies (or other templates) may be hypothesized to play a causal role. If composers provide a causal link from perception to the repertoire, the templates are, themselves, a function of bottom-up processes.

Figure 5.3: The causal paths implied by different types of model. In both cases, measurable variables are in the top row, latent mental processes (which we may attempt to model) are in the middle row, and the resulting perception (e.g., 'degree of fit', which may be reported as empirical data) is in the bottom row. Note that *top-down* and *bottom-up* do not refer to the direction taken by the arrows (causal paths); they refer to whether or not *previous music events* is an input variable to the perceptual process.

thought of as a reversal of that used to model the probe tone data. For the probe tone data, I modelled the fit of the probes by calculating their spectral pitch class similarity to a weighted major or minor triad that represents the tonal centre (and was also the final chord played in all of the contexts). To reverse this, I start with a 'context' scale and model the fit of a variety of 'probe' pitch classes and 'probe' major and minor chords to this scale.

I hypothesize that scale pitches with low fit may be heard as more musically tense or active (in need of resolution) than scale pitches with high fit (which may serve as melodic tonics), and major and minor chords with the highest fits may tend to function as that scale's tonic triad. Parncutt uses this technique to determine the pitch class tonics

for the diatonic scale (Parncutt and Prem, 2008; Parncutt, 2011) in medieval music, though his approach is inconsistent in the same way as in the Parncutt 88/11*a* model (he measures the commonality of the context scale pitches and the probes but, while the contextual scale pitches include virtual pitch classes, the candidate tonic probes do not). In the following examples, I additionally look for tonic triads as well as pitch classes—I use the spectral pitch class similarity of contextual scale pitches and probe pitches, or probe triads, to model their tonicness. Furthermore, I model the context and probes consistently—their pitches have identical harmonic spectra. This allows us to explore the profiles of fit produced by a variety of scales. In every case, I used $\rho = 0.67$ and $\sigma = 5.95$, as optimized for model c (I could have chosen the values as optimized for any of the three models, but model c's values fall between those of models a and b, so seemed a sensible choice; anyway, the results are robust over the three sets of values).

For this 'reverse' model to make sense requires that we consider the scales as known entities (in either short-term or long-term memory). For a scale to be known, it must be perceived as a distinct selection of pitches or as a specific subset of a chromatic gamut of pitch classes. A composer or performer aids this by ensuring all scale pitch classes are played over a stretch of time short enough for them all to be maintained in short-term memory, and by utilizing scales that have relatively simple and regular structures such as well-formed scales.[9] Long-term memory is also likely to play an important role in that certain scales are learned through repetitive exposure.

In this chapter, I have previously used uppercase Roman numeral notation, so IV–V–I in a major key means all chords are major, while IV–V–I in a minor key means the first and last chords are minor.[10] In the following sections I am dealing with specific scales, so I use upper

9 Well-formed scales are defined in Sec. 3.6.2. They comprise just two step sizes, evenly distributed, and are exemplified by the diatonic and pentatonic scales. They can also be generalized to microtonal forms with a variety of structures (Erlich, 2006; Milne et al., 2011a).

10 In a minor key, the V chord is taken to be major because this is much more common than a minor V.

case to denote major triads and lower case to denote minor. For example, the above minor tonality cadence is now denoted iv–V–i.

5.3.1 *Fit Profiles for 12-TET Scales*

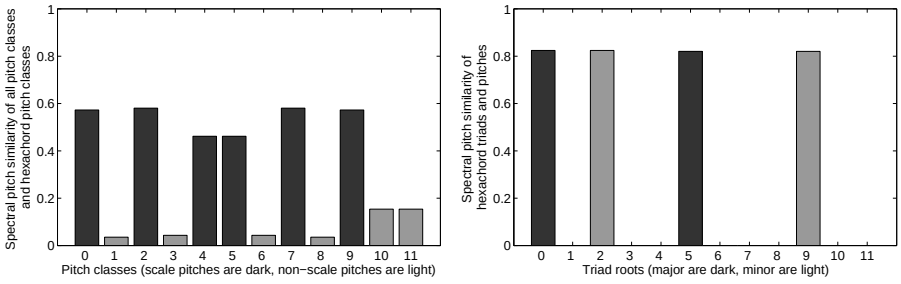
In this section, I consider a variety of scales that can be thought of as subsets of the twelve pitch classes of twelve-tone equal temperament. This may seem anachronistic given that many of the scales discussed in this subsection probably had origins prior to the widespread usage of 12-TET. However, there is little evidence to suggest that relatively small differences in tuning (e.g., between the meantone tunings prevalent at the birth of tonality and the strict twelve-tone equal tunings of much contemporary auto-tuned popular music) has a strong influence on basic tonal effects; for example, the strong closure induced by a IV–V^7–I cadence seems invariant across these settings.[11] Furthermore, I demonstrate in Section 5.3.2 that my model produces similar results for the diatonic scale tuned to 12-TET and to quarter-comma meantone. In this latter section, I also explore some interesting tonal effects produced when the tuning, and resulting scales, are radically different to those obtained in 12-TET.

5.3.1.1 *Major (Guidonian) hexachord.*

This six-tone scale formed the basis of much medieval music theory and pedagogy (Berger, 1987). It is equivalent to a diatonic scale with the fourth or seventh scale degree missing. For instance the C hexachord contains the pitches C, D, E, F, G, A. There is no B or B♭ to fill the gap between A and C. In modal music, the note used to fill the gap was either a *hard* B (a B♮) or a *soft* B (a B♭).[12] The choice of hard or soft was not notated but was made by performers to avoid simultaneous or melodic tritones—this practice is called *musica ficta* (Berger, 1987).

11 Although basic tonal effects may be invariant, equal temperaments do provide certain advantages because they facilitate unlimited modulation, and enharmonic substitution and modulation. In that sense, they open up greater compositional resources.

12 The shape of the natural and flat symbols derive from two different ways of writing the letter 'b'.



(a) Modelled fit of major hexachord pitches.

(b) Modelled fit of major hexachord chords.

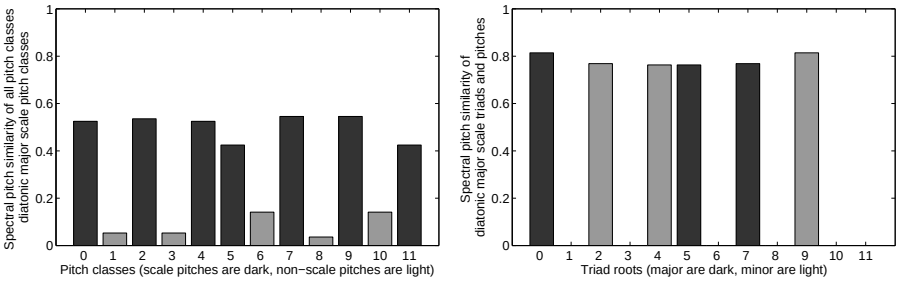
Figure 5.4: Modelled pitch class and chord fit with the Guidonian hexachord.

In Figure 5.4, I will assume that pitch class 0 corresponds to C. Figure 5.4a shows that the pitch classes E and F (4 and 5), which are a semitone apart, are the least well-fitting of the hexachord tones. In Gregorian chant, the *finalis* (final pitch) was D, E, F, or G (corresponding to the modes *protus*, *deuterus*, *tritus*, and *tetrardus*). Of these modes, Figure 5.4a shows that the pitch classes with the highest fit are at D and G (2 and 7), which suggests these two modes have the most stable final pitches. This tallies with statistical surveys, referenced in Parncutt (2011), which indicate these two modes are the most prevalent. (The relative fits of D and G are even higher when the hexachord has a Pythagorean tuning in which all its fifths have the frequency ratio 3/2—such tunings were prevalent prior to the fifteenth century (Lindley, 2013).)

When we look at the modelled fit of each of the hexachord's major and minor triads with all the pitches in the hexachord, the results are quite different (Figure 5.4b). Here, every major or minor chord has identical fit with this scale. It is as if the Guidonian hexachord—when used for major/minor triad harmony—has no identifiable best-fitting tonic chord. As shown in the next example, all of this changes when that missing seventh degree is specified, thereby producing a specific diatonic scale.

5.3.1.2 *Diatonic major scale.*

The diatonic scale—regardless of its mode—has numerous properties that make it perceptually and musically useful. A number of those



(a) Modelled fit of major scale pitches. (b) Modelled fit of major scale chords.

Figure 5.5: Modelled pitch class and chord fit with the major scale..

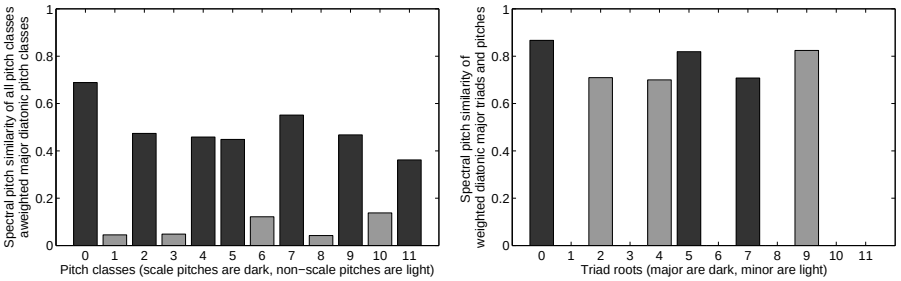
properties follow from its well-formedness (Wilson, 1975; Carey and Clampitt, 1989) such as Myhill's property, maximal evenness, uniqueness, coherence, and transpositional simplicity (space precludes explaining these properties here—they are summarized in Prechtl et al. (2012)). Furthermore, it contains numerous consonant intervals (approximations of low integer frequency ratios), and supports a major or minor triad on all but one of its scale degrees. For tonal-harmonic music, the major scale (e.g., C, D, E, F, G, A, B) is the most important and prevalent mode of the diatonic scale. The only other mode that comes close is the Aeolian (e.g., A, B, C, D, E, F, G, or C, D, E♭, F, G, A♭, B♭)—also known as the *natural minor scale*—which is one of the three scale forms associated with the minor scale (the other two are the harmonic minor, in which the Aeolian's seventh degree is sharpened, and the ascending melodic minor in which the sixth and seventh degrees are sharpened).

The addition of a seventh tone to the hexachord—thereby making a diatonic scale—makes the fits of its triads more heterogeneous. Figure 5.5b illustrates this with the diatonic major scale—note how the Ionian and Aeolian tonics (roots on pitch classes 0 and 9, respectively) are modelled as having greater fit than all the remaining triads. This, correctly, suggests they are the most appropriate tonics of the diatonic scale—the major scale's tonic and the natural minor scale's tonic, respectively. This is also reflected in the common usage of the submediant chord (vi) as a substitute for the tonic (I) in deceptive cadences (Piston and Devoto 1987, p. 191; Macpherson 1920, p. 106), and the

frequent modulation of minor keys to their relative major (Piston and Devoto, 1987, p. 61). It is also interesting to observe that the fourth and seventh degrees of the major scale have lower fit than the remaining tones. This possibly explains why these two scale degrees function as leading tones in tonal-harmonic music—scale degree $\hat{7}$ resolving to $\hat{1}$, and $\hat{4}$ resolving to $\hat{3}$—for example, both these motions occur in the dominant seventh to tonic cadence (i.e., V^7–I).

There are five aspects of major-minor tonality not obviously explained by the above fit profiles: (a) in the diatonic scale, the Ionian tonic is privileged over the Aeolian tonic; (b) in the major scale, the seventh scale degree is typically heard as more active—more in need of resolution—than the fourth degree; (c) the importance of the V–I cadence; (d) the activity of the seventh degree of the major scale is significantly reduced when it is the fifth of the iii (mediant) chord in comparison to when it is the third of the V (dominant) chord. I propose two additional hypotheses that may account for these features.

A hypothesis to explain the first two features is that the strongest sense of harmonic resolution is induced when a bad-fitting tone moves by semitone to the *root* of a best-fitting chord. In the white-note diatonic scale, there are two best-fitting triads (Cmaj and Amin) and two worst-fitting pitch classes (B and F). This means that only Cmaj has a root (C) that can be approached by semitone from a worst-fitting pitch class (B); for Amin, the root (A) cannot be approached, by semitone, by either B or F. If we assume that this provides an in-built advantage to the Ionian mode, this introduces an interesting feedback effect. Let us now weight the pitch class C a little higher than the other tones to reflect its status as the root of a best-fitting triad that is approached, by semitone, by a worst-fitting pitch—the results of this are illustrated in Figure 5.6 where the weight of C is twice that of the other tones (possibly an extreme value, but it demonstrates the effect). Although the pitch class C is a member of both the C major and A minor tonics, Figure 5.6b shows that increasing its weight disproportionately enhances the fit of the triad Cmaj over the triad Amin. It also decreases the fit of



(a) Modelled fit of major scale pitches, when tonic pitch class is double-weighted.

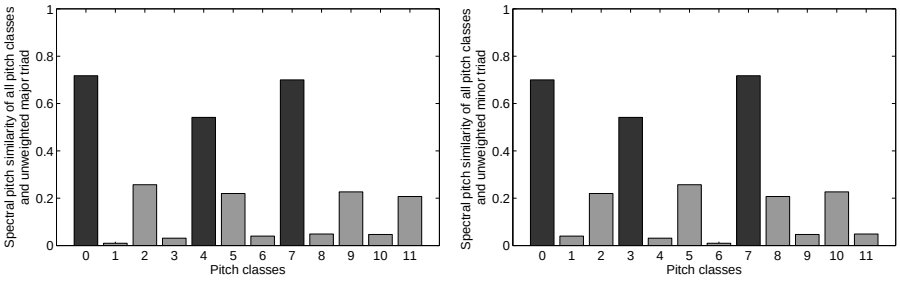
(b) Modelled fit of major scale chords, when tonic pitch class is double-weighted.

Figure 5.6: Modelled pitch class and chord fit with a major scale with a double-weighted tonic pitch class.

B (Figure 5.6a). It seems likely, therefore, that this results in a positive feedback loop: I hypothesize that the resolution of the poor-fitting B to the root of Cmaj increases the perceived fit of C; I model this by giving the C a greater weight, and this disproportionately increases the fit of Cmaj over Amin, and reduces the fit of B; this is likely to result in an even stronger resolution from B to the root of Cmaj (B is worse fitting than before, and Cmaj is better fitting) and this, in turn, will further enhance the fit of pitch class C and thereby enhance the fit of Cmaj over Amin, and so on in a positive feedback loop.

The third feature—the importance of the V–I cadence, which is typically described as the 'strongest' or 'most powerful' progression in tonal music (Piston and Devoto 1987, p. 21; Pratt 1996, p. 9)—also follows, in part, from the same hypothesis that resolution is enhanced by a low-fit pitch moving to the root of a high-fit triad. This favours the resolutions V–I or vii°–I (which contain the scale degrees $\widehat{7}$–$\widehat{1}$—a resolution to the tonic's root), over IV–I or ii–I (which contain the scale degrees $\widehat{4}$–$\widehat{3}$—a resolution to the tonic's third). It is also interesting to note that V^7–I and vii°–I, which provide the strongest tonal resolutions, contain both $\widehat{7}$–$\widehat{1}$ and $\widehat{4}$–$\widehat{3}$.

However, this suggests that iii–I would also provide an effective cadence because it too has the worst-fitting $\widehat{7}$ resolving to the root of I. But such cadences are rare (Piston and Devoto, 1987, p. 21), and the activity of the seventh degree is typically felt to be much reduced when it



(a) Unweighted major triad. (b) Unweighted minor triads.

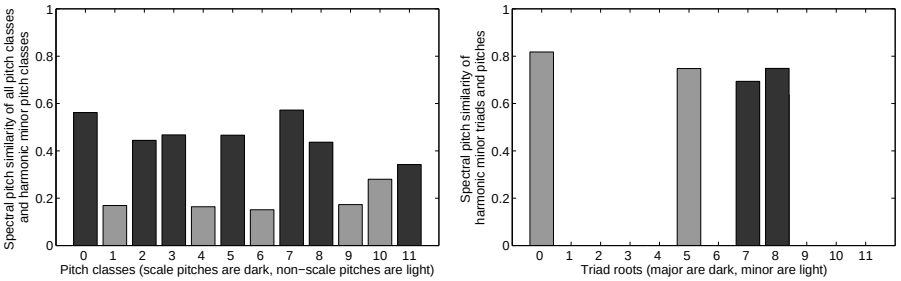
Figure 5.7: Modelled pitch class fits with unweighted major and minor triads.

is the fifth of the iii chord—a common use of the iii chord is to harmonize the seventh degree when it it is descending to the sixth (Macpherson, 1920, p. 113). This may be explained by a second hypothesis, which is that we need to consider the fit of pitches not just in relation to their scalic context, but also in relation to their local harmonic (chordal) context. Against the context of a major or minor chord, the third is the worst-fitting pitch—see Figure 5.7 (all triad pitches are equally weighted), which shows that both chords' thirds (pitch class 4 for the major triad, and 3 for the minor) have lower fit than the root and fifth (pitch classes 0 and 7). This suggests that the higher fit of scale degree $\widehat{7}$ in iii—due to it being the chord's fifth—makes it less active; while the lower fit of $\widehat{7}$ in V—due to it being the chord's third—makes it more active. This hypothesis, therefore, explains the greater stability of the seventh degree in iii compared to V, and completes the explanation for the importance of the V–I, V^7–I, and vii°–I cadences.

These additional hypotheses (the importance of semitone resolutions from poor-fit tones to roots of good-fit triads, and the decreased fit of pitches that are the thirds of chords) seem promising; in future work, I hope to precisely specify these effects, and use them to model responses to a variety of chord progressions and scalic contexts.

5.3.1.3 *Harmonic minor scale.*

An important aspect of the minor tonality is that the harmonic minor scale is favoured over the diatonic natural minor scale—particularly in common practice cadences where (the harmonic minor) V–i is nearly



(a) Modelled fit of harmonic minor scale pitches.

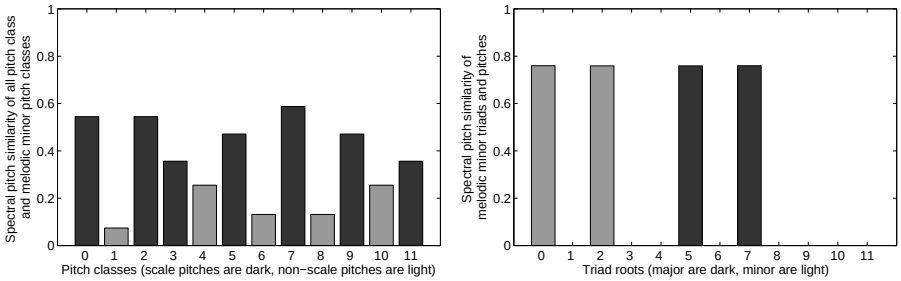
(b) Modelled fit of harmonic minor scale chords.

Figure 5.8: Modelled pitch class and chord fit with the harmonic minor scale.

always used in preference to (natural minor) v–i (Piston and Devoto, 1987, p. 39). The harmonic minor scale is equivalent to the Aeolian mode with a sharpened seventh degree. This change has an important effect on the balance of chordal fits—and goes some way to explaining why this scale forms the basis of minor tonality in Western music.

Figure 5.8a shows that $\widehat{7}$ is clearly the worst-fitting scale degree; the next worst are $\flat\widehat{6}$ and $\widehat{2}$. Figure 5.8b shows that the best-fitting triad is i; furthermore, every pitch in this tonic i chord can be approached by the three most poorly-fitting scale degrees which, therefore, act as effective leading tones: $\widehat{7}$–$\widehat{1}$, $\flat\widehat{6}$–$\widehat{5}$, and $\widehat{2}$–$\flat\widehat{3}$—as exemplified by a chord progression like Bdim7–Cmin, or G7♭9–Cmin. These properties appear to make this scale a context that provides unambiguous support of a minor triad tonic. Compare this to the diatonic mode, where there is an equally well-fitting major triad; for example, Macpherson (1920, p. 162) writes that

> any chord containing the minor 7th usually requires to be followed as soon as possible by a chord containing the Leading-note … otherwise the tonality easily becomes vague and indeterminate, and the music may tend to hover somewhat aimlessly between the minor key and its so-called ‘relative’ major.



(a) Modelled fit of melodic minor scale pitches.

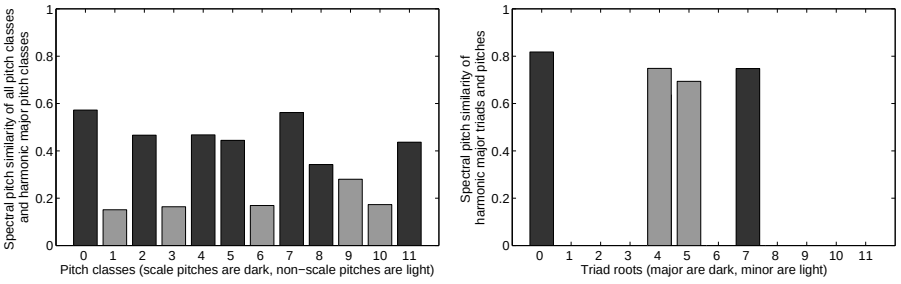
(b) Modelled fit of melodic minor scale chords.

Figure 5.9: Modelled pitch class and chord fit with the ascending melodic minor scale.

5.3.1.4 *Ascending melodic minor scale.*

It is well-recognized in music theory that the harmonic minor scale provides effective harmonic support for a minor tonic, but that it is also melodically awkward due to the augmented second between its sixth and seventh degrees. When a melodic line is moving from the sixth to the seventh degree, this awkward interval is typically circumvented by sharpening the sixth degree—this produces the ascending melodic minor scale (the descending melodic minor scale is identical to the natural minor scale (Aeolian mode)).

Figure 5.9b shows that, in terms of chord fits, this scale has returned to a similar situation as the Guidonian hexachord: all chords have equal fit, hence there is no obvious tonic. This suggests that using this scale, for brief periods of time to improve the melodic line, will not disrupt a minor tonality previously established with the parallel harmonic minor scale. However, this scale cannot form the foundation of a minor tonality, because it has no specific tonal centre (when triads are used). Again, this seems to be in accord with conventional tonal music theory, which specifies that the primary function of this scale is to improve melodic connections rather than to provide the basis for harmony (the use of the raised sixth degree is usually subject to strict melodic conventions—e.g., Schoenberg (1969, p. 18) advises that it should not move to the natural sixth or natural seventh degrees).



(a) Modelled fit of harmonic major scale pitches.

(b) Modelled fit of harmonic major scale chords.

Figure 5.10: Modelled pitch class and chord fit with the harmonic major scale.

5.3.1.5 *Harmonic major scale*.

In the same way that sharpening the seventh degree of the Aeolian mode can make its tonic unambiguously the best-fitting, it is interesting to consider if there is a different alteration that can do the same for the Ionian mode. The alteration that seems to provide a similar benefit for the Ionian is to flatten its sixth degree, to make the harmonic major scale.

In comparison to Figure 5.5b, Figure 5.10b shows how the I chord is now the uniquely best-fitting chord. This appears to indicate that flattening the sixth degree of the major scale strengthens the major tonality. This accords with Harrison's description of the chromatic iv in major as the tonic-strengthening dual of the 'chromatic' V in minor (1994, pp. 15–34). However, like the harmonic minor scale, this alteration creates an awkward sounding melodic interval—the augmented second between the sixth and seventh degrees—which maybe explains why this scale has not become the default major tonality scale.

5.3.2 *Fit Profiles for Microtonal Scales*

Unlike all of the previously discussed models, mine is generalizable to pitches with any tuning (e.g., microtonal chords and scales). It is interesting to explore some of the predictions of pitch class and chord fit made by the model given a variety of microtonal scales. All of the microtonal scales I analyse here are well-formed. I do this under the hy-

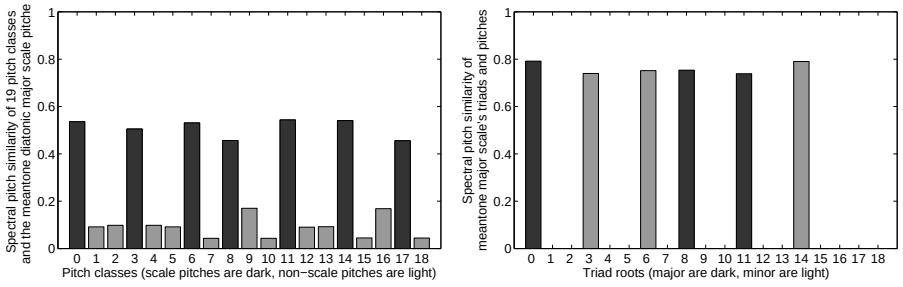
pothesis that the simple and regular structure of such scales may make them easier to hold in short-term memory, or learn as part of long-term memory.[13]

5.3.2.1 *Quarter-comma meantone diatonic scale.*

This tuning was first described by Pietro Aaron in 1523 (cited in Barbour (1951)) who described a system of temperament where every perfect fifth is equally flattened slightly but all major thirds are perfectly tuned. This is around the time that modal music began its gradual transition into harmonic tonality, and may have been a prevalent tuning at that time. For that reason it is interesting to see what, if any, impact it has on the fit of the diatonic pitches and chords. One aspect that differentiates meantone tunings from 12-TET is that enharmonically equivalent pitches (e.g., C♯ and D♭) do not have identical tunings. For this reason, I use a gamut of 19 pitch classes (e.g., the chain-of-fifths from C♭ to E♯), which provides a sharp and a flat for every diatonic scale degree (e.g., C, D, E, F, G, A, B) except for the fourth (e.g., F) which has no flat, and the seventh (e.g., B) which has no sharp. Another difference is that its major and minor triads are, by any standard metric, closer to the low integer ratios of just intonation (4:5:6 and 10:12:15, respectively) than the 12-TET versions: the just intonation triads are, to the nearest cent, $(0, 386, 702)$ and $(0, 316, 702)$; the 1/4-comma meantone triads, to the nearest cent, are $(0, 386, 697)$ and $(0, 310, 697)$; the 12-TET triads are $(0, 400, 700)$ and $(0, 300, 700)$.

For the diatonic scale degrees and chords, the overall pattern of fits is similar to that produced by 12-TET—as shown in Figure 5.11. The fourth and seventh scale degrees are still modelled as the worst fitting, and the Ionian and Aeolian tonic triads are still modelled as the best fitting. This suggests that this pattern and, hence, its tonal implications,

13 Equal step scales are structurally simpler and more regular than well-formed scales, but they are actually too regular because their internal structure is completely uniform—every pitch class or chord bears the same relationship to all other scale pitches and chords. The structure of equal step scales cannot, therefore, support a different musical function on different scale degrees—such a musical function may be imposed by pitch repetition or a drone, but it is not inherent to the scale, merely to its usage.



(a) Modelled fit of 1/4-comma diatonic major scale pitches.

(b) Modelled fit of 1/4-comma diatonic major scale chords.

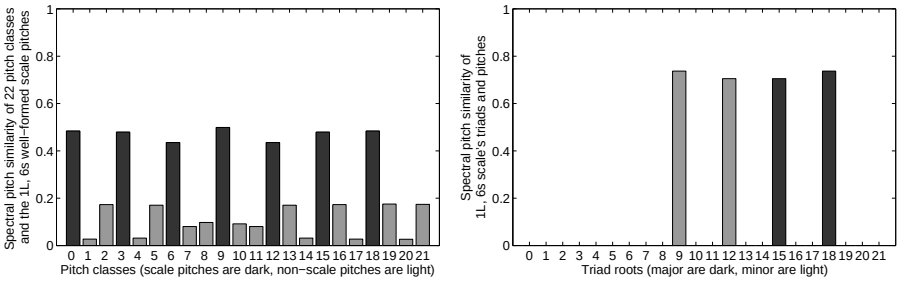
Figure 5.11: Modelled pitch class and chord fit with the 1/4-comma meantone diatonic major scale.

are robust over such changes in the underlying tuning of the diatonic scale.

5.3.2.2 *22-TET 1L, 6s porcupine scale.*

In the following three examples, I look at different well-formed scales that are subsets of 22-tone equal temperament. The names of these temperaments (*porcupine*, *srutal*, and *magic*) are commonly used in the microtonal community, and are explained in greater detail in Erlich (2006) and the website `http://xenharmonic.wikispaces.com/`. In all of these scales, the tunings—rounded to the nearest cent—of the major triads are $(0, 382, 709)$, and the tunings of the minor triads are $(0, 327, 709)$. These tunings are, by most standard metrics, closer to the just intonation major and minor triads than those in 12-TET. For each scale, the spectral pitch class similarities suggest one or more triads that will function as tonics. I do not, at this stage, present any empirical data to substantiate or contradict these claims; but I suggest that collecting such empirical data—tonal responses to microtonal scales—will be a useful method for testing bottom-up models of tonality. Audio examples of the scales, their chords, and some of the cadences described below, can be downloaded from `www.dynamictonality.com/probe_tone_files/`.

The porcupine scale has seven tones and is well-formed—it contains one large step of size 218 cents and six small steps of size 164 cents (hence its signature 1L, 6s), and the scale pitch classes are indicated with dark bars in Figure 5.12a. Figure 5.12b shows that the major triad on 18



(a) Modelled fit of porcupine 1L, 6s scale pitches.

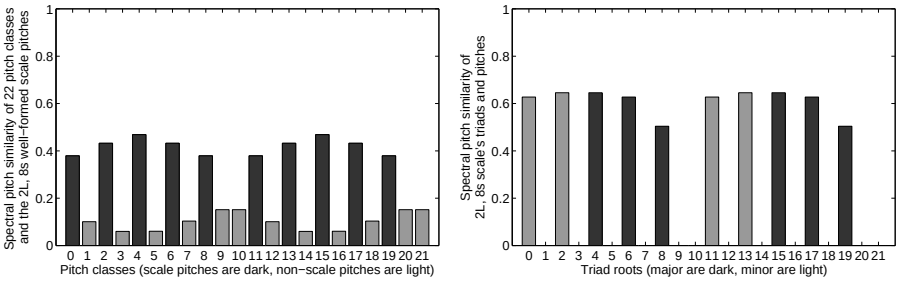
(b) Modelled fit of porcupine 1L, 6s scale chords.

Figure 5.12: Modelled pitch class and chord fit with the porcupine 1L, 6s scale.

and the minor triad on 9 are modelled as the best-fitting. This suggests that, within the constraints of this scale, they may function as tonics. The worst-fitting pitch classes are 6 and 12, which can both lead to the root of the minor triad on 9. Neither of these potential leading tones are thirds of any triads in this scale, which possibly reduces their effectiveness when using triadic harmony. However, the above suggests the most effective cadences in this scale will be the minor chord on 12 leading to the minor chord on 9, the major chord on 15 (whose fifth is pitch class 6) leading to the minor chord on 9, or a variety of seventh chords containing both 6 and 12 like the dominant seventh built on 15 (whose third is 6 and seventh is 12) also leading to the minor chord on 9. Using Roman numerals, taken relative to the minor tonic on pitch class 9, these are ii–i, III–i, and III^7–i, respectively.

5.3.2.3 *22-TET 2L, 8s srutal scale.*

This ten-tone microtonal scale—first suggested by Erlich (1998)—is unusual in that it repeats at the half-octave (it is well-formed within this half-octave interval). This repetition accounts for why the fit levels—shown in Figure 5.13—also repeat at each half-octave. It contains two large steps of size 164 cents, and eight small steps of size 109 cents. The scale pitches are indicated with dark bars in Figure 5.13a. The modelled fits suggest there are two possible major triad tonics (on pitch classes 4 and 15) and two possible minor tonics (on pitch classes 2 and 13). The roots of both the minor chords can be approached by a poorer-fitting



(a) Modelled fit of srutal 2L, 8s scale pitches.

(b) Modelled fit of srutal 2L, 8s scale chords.

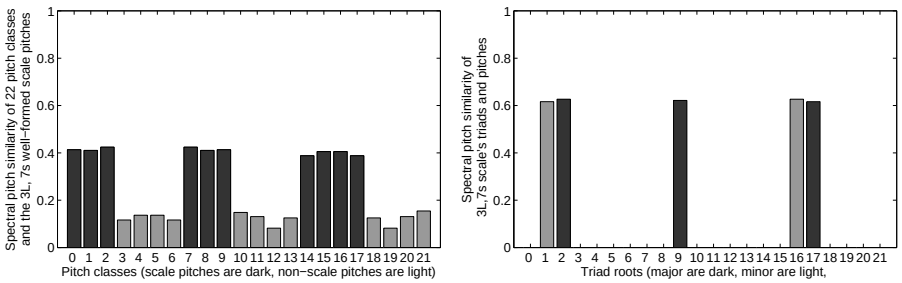
Figure 5.13: Modelled pitch class and chord fit with the srutal 2L, 8s scale.

leading tone (pitch classes 0 and 11) than can the major (pitch classes 2, 6, 13, and 17). This suggests effective cadences can be formed with the major chord on 15 (whose third is pitch class 0) proceeding to the minor chord on 2 (or their analogous progressions a half-octave higher), or variety of seventh chords such as the dominant seventh on 4 (whose seventh is pitch class 0). Using Roman numerals relative to the minor tonic on 2 (or 13), these are VII–i and II^7–i, respectively. These cadences can be thought of as slightly different tunings of the familiar 12-TET progressions V–i and $\flat\text{II}^7$–i.

5.3.2.4 *22-TET 3L, 7s magic scale.*

This microtonal scale also has ten tones, and is well-formed with respect to the octave (so no repetition at the half-octave)—it has three large steps of size 273 cents and seven small steps of size 55 cents. As before, the dark bars in Figure 5.14a indicate the scale pitches. In this scale, every degree that is a root of a major triad is also a root of a minor triad (and vice versa). For this reason, in Figure 5.14b, only the better fitting (major or minor) is shown on the chart; for the pitch class 9, however, the major and minor triad have equal fit, so this should be borne in mind.

The modelled fits, in Figure 5.14b, suggest two possible major tonics (with roots on pitch classes 2 and 9) and two possible minor tonics (on pitch classes 9 and 16). Figure 5.14a shows that, in terms of fit, pitch class 17 looks like a promising leading tone to the root of the minor



(a) Modelled fit of magic 3L, 7s scale pitches.

(b) Modelled fit of magic 3L, 7s scale chords.

Figure 5.14: Modelled pitch class and chord fit with the magic 3L, 7s scale.

triad on 16. However, this pitch class is not the third of any triad in the scale. The other leading tone contenders are on 1 and 8, and both of these can be triad thirds. This implies the major chord on 2, and the major or minor chord on 9, may function as tonics in this scale. This suggests effective cadences can be formed with the major chord on 16 (whose third is pitch class 1) proceeding to the major triad on pitch class 2, or the major chord on pitch class 1 (whose third is pitch class 8) proceeding to the major or minor triad on pitch class 9. In Roman numeral notation, relative to their respective tonics, these are VII–I, VII–I, and VII–i. Interestingly, in all these examples the cadences are—in terms of 12-TET—similar to a major chord, whose root is pitched in-between V and ♭VI, proceeding to I or i (the distance between these roots is 764 cents).

5.4 CONCLUSION

In this chapter, I have shown that there at least two types of plausible bottom-up model—Parncutt's virtual pitch class commonality models, and my spectral pitch class similarity models—that can explain why the probe tone data take the form they do. I argue that bottom-up explanations, such as these, are able to account not just for the existence of fit profiles (as provided by top-down models), but also for the specific form they take. In light of both theories' ability to explain and predict the data, I suggest that there is now little reason to believe the probe

tone data are a function purely of top-down processes. I cannot, on the basis of the probe tone data, determine whether the primary mechanism is spectral pitch class or virtual pitch class similarity. To distinguish between these effects would require novel experiments.

I have also used my model in the reverse direction—to predict candidate tonic triads for a number of scales that are subsets of the full twelve chromatic pitch classes. The results accord well with music theory. Furthermore, I have also suggested some additional mechanisms that may account for strong cadences (a poor-fitting tone moving to the root of a best-fitting triad) and how this, in turn, may cause the diatonic scale to become more oriented to its major (Ionian) tonic rather than its minor (Aeolian) tonic. I also suggest a possible reason for why the seventh degree loses much of its activity (need to resolve) when it is the fifth of the mediant (iii) chord. And, in combination, these mechanisms support the use of V–I as a cadential chord progression. These latter hypotheses are somewhat speculative because they have not been included in a formal mathematical model, but I feel they are promising ideas that warrant further investigation.

I have also claimed my model can challenge the notion that there is a tonal hierarchy, which is an unchanging (or slowly evolving) template against which recently heard pitches are compared. Rather, in my template-independent theory, I suggest that any given musical context automatically generates a corresponding profile of fits. For certain commonly-used scales, such profiles may become embedded as tonal hierarchies (templates), as might the tonal implications of common melodic lines and harmonic progressions. But crucially, under my theory, the templates are not the initial cause of tonal functionality; rather, they are an effect of the more basic and universal psychoacoustic process of spectral (or virtual) pitch similarity. This implies that our cognition of tonality, and the types of tonal musics we create are—to some extent—constrained and shaped, in a nontrivial and predictable way, by our perceptual and cognitive apparatus.

Finally, I have pointed to the way in which microtonal scales can also be analysed with this technique, and how this may become an important means to explore our general perception of tonality, and to test models thereof. Ideally, any model that purports to explain—from the bottom up—how Western tonality works, should also be able to make useful predictions for the possibly different tonalities evoked by completely different scales and tunings.

6

CONCLUSION

In this dissertation, my aim has been to identify and model the innate processes by which feelings of tension, resolution, stability, and so forth, are induced by successions of pitches and chords, irrespective of their harmonic consonance.

I have chosen to focus on innate mental processes—those aspects of music perception that are not due to long-term familiarity with a specific corpus of music. This is because only such processes can explain (specify the casual origins of) associations between acoustical events and mental phenomena.

In order to do this, I have postulated that—given a context of pitches, such as a scale—those chords that are the most consonant, have the greatest affinity, and are the most familiar will tend to be heard as the most stable and tonic-like. Familiarity is, by definition, a top-down process, but both consonance and affinity have plausible bottom-up psychoacoustic models. I have focused my attention on bottom-up models of affinity, because it is clear that consonance can provide only a partial answer (e.g., in the context of a C major scale, the root-position triads Cmaj and Gmaj have identical consonance but differing levels of stability).

When an instrument plays a single notated pitch, it actually produces a multiplicity of spectral (and possibly virtual) pitches. Following Terhardt and Parncutt, I have hypothesized that the affinity of any two tones or chords is due, in part, to the similarity of their spectral or virtual pitches. In order to effectively model such similarities, I have developed a novel family of representations of pitches called expectation tensors.

In Chapter 3, I demonstrated how expectation tensors can model the uncertainties of pitch perception by 'smearing' each pitch over a range

of possible values, and the width of the smearing can be related to experimentally determined frequency difference limens. The tensors can embed either absolute or relative pitches: in the latter case, embeddings of pitch collections that differ only by transposition have zero distance; a useful feature that relates similarity to structure. Furthermore, tensors of any order (dimensionality) can be formed, allowing the embeddings to reflect the (absolute or relative) monad (pitch), dyad (interval), triad, tetrad, and so forth, content of the pitch collection. The distance between expectation tensors of the same order can be determined with any standard metric or similarity measure (such as L_p or cosine).

I also demonstrated how absolute monad expectation tensors can be used to embed the spectral pitches (or pitch classes) of tones or chords, and how the cosine similarity of such vectors can be used to model the affinity of the tones or chords they embed. In Chapter 4, I described an experiment that eliminates the confounding top-down influence of horizontal familiarity (that part of affinity that is a function of each interval's prevalence). In the absence of this confound, the data indicate that spectral pitch similarity is an effective bottom-up model of affinity. In other words, there is a psychoacoustic component to listeners' perception of the extent to which tones with different pitches fit together—the greater the similarity of their spectra, the greater their affinity. The data also show that a spectral pitch based model of harmonicity (toneness) is also correlated with perceived affinity but, in this case, this may be modelling either an innate or a learned process.

I additionally showed how these results indicate that certain common scales—such as those based on tetrachords or generated by perfect fifths like the pentatonic and diatonic—maximize the affinity of the harmonic complex tones produced by the human voice and most pitched musical instruments. Such scales may, therefore, be a natural consequence of our perceptual apparatus, irrespective of culture. I also suggest that synthetic sounds with timbres matched to the underlying tuning can be used not just to maximize consonance, but also to maximize melodic affinity for microtonal scales.

In Chapter 5, I showed that spectral pitch class similarity can also model Krumhansl's probe tone data with considerable accuracy. In light of the ability of this model, and Parncutt's virtual pitch model, to both explain and predict the data, I suggest there is now little support for the widely held belief that tonal perceptions are due only to top-down processes. In other words, I have provided compelling evidence that bottom-up processes play an important role in the perception of tonal fits. This implies there is no requirement to interpret the probe tone data as representing a long-term memory template (a tonal hierarchy). I do not argue such templates do not exist—it seems highly plausible that learned processes do play a meaningful role. But, importantly, bottom-up explanations mean the initial causal impetus for their forms comes from psychoacoustically based processes. In other words, our cognition of tonality, and the types of tonal musics we create are constrained and shaped, in a nontrivial and predictable way, by our perceptual and cognitive apparatus.

I also used the spectral pitch class similarity model to predict, for a variety of scales, which triads are likely to function as tonics, and which pitches are likely to function as tense leading tones. The model's predictions concur with music theory. I additionally suggested three mechanisms that may account for why: the diatonic scale is biased towards its major (Ionian) mode rather than its natural minor (Aeolian) mode; the seventh degree of the major scale loses much of its tension when it is the fifth of the iii chord; the V–I progression is so important in cadences. These mechanisms are that (a) given a scale, the strongest sense of harmonic resolution is induced when a bad-fitting tone moves by a small interval (e.g., a semitone) to the root of a best-fitting chord; (b) there is a feedback mechanism whereby the increased salience of a tonic degree affects the instability of its leading tone(s); (c) the tension of pitches is a function of their affinity with their local harmonic context (chord degree) in addition to their scalic context (scale degree).

At this stage, these latter hypotheses are somewhat speculative because they have not been included in a formal mathematical model.

Finally, I have pointed to the way in which microtonal scales can also be analysed with this same technique.

In summary, I have developed a novel set of techniques to model the perceived similarity of pitch collections, and I have used them to build bottom-up models for certain important aspects of tonal cognition—affinity and tonal stability. Experimental tests of the models have shown them to be effective at both predicting and explaining these aspects of tonality.

6.1 CONTRIBUTIONS

The first novel contribution provided in this dissertation is my fourfold categorization of mental processes relevant to music cognition: extrinsic nurture, intrinsic nurture, extrinsic nature, and intrinsic nature. This categorization is related to that provided by Sloboda and Juslin (2001), but differs in that they do not explicitly separate intrinsic processes into nature and nurture. I additionally illustrate how these categories are related to those used in semiotic theory and those suggested by Juslin and Västfjäll (2008). The principal purpose of my categorization is to enable a clear and unambiguous distinction to be made between top-down and bottom-up models: the former require, as an input variable, a statistical analysis of a musical corpus; the latter do not.

The second novel contribution is the development of expectation tensors. Prior to expectation tensors, there had been no generalized method to represent collections of pitches, intervals, triads, and so forth within a principled probabilistic framework and incorporating basic psychoacoustic processes of pitch perception. Furthermore, by generalizing the resulting embeddings into multi-dimensional forms (i.e., tensors) and allowing for pitches to be represented in either absolute or relative form, I have constructed a family of pitch embeddings that are generalizations of a number of familiar embeddings used in musical set theory (i.e., interval vectors and other subset-class vectors).

The third novel contribution has been the development of an expectation tensor (spectral pitch vector) model of tonal affinity. This model has a small number of parameters (spectral roll-off ρ and smoothing width σ). It has some similarities to Parncutt's virtual pitch commonality model, but it differs in that: it uses spectral rather than virtual pitches; it does not assume each pitch is categorized as a chromatic pitch class regardless of its precise tuning; it inherits the principled psychoacoustic and probabilistic foundations of the expectation tensors.

The fourth novel contribution has been utilizing microtonal stimuli to experimentally disambiguate innate and learned processes (as defined in this dissertation). The use of microtonal stimuli in experimental investigations of music perception is rare; I am aware of only a few researchers who have used microtonal stimuli—Vos (1982, 1984, 1986, 1988); Vos and van Vianen (1985b,a), Bucht and Huovinen (2004), and Ferrer (2007)—and in none of these cases are the microtonal stimuli used to disentangle nature and nurture (as defined in this dissertation).

The fifth novel contribution has been to demonstrate that the perceived fit of successive pitches (affinity) is a causal function of their spectral pitch similarity—a model of an innate mental process.

The sixth novel contribution has been to show that affinity is also a function of the harmonicity of the timbres used. It is not possible to say whether this is due to participants' familiarity with harmonic complex tones, or whether it is modelling an innate preference for such timbres.

The seventh novel contribution has been to demonstrate unequivocally that a bottom-up psychoacoustic model can account for Krumhansl's probe tone data. And that the same model also makes realistic predictions about the musical functions of chords in a variety of familiar scales.

The eighth novel contribution is that I have applied my model to microtonal scales. As far as I know, this is the only model able to make predictions about the perceived tonal effects (affinities, stabilities, ten-

sions) of successively played pitches and chords within such scales.[1] At this time, I have no empirical evidence to support these predictions.

6.2 LIMITATIONS

None of the experiments conducted for this dissertation have enabled me to determine the relative importance of spectral and virtual pitches to tonal perception. However, it seems that models containing only spectral pitches (or pitch classes) are highly effective. The only way to distinguish the effect sizes of spectral and virtual pitches is to conduct experiments where their predictions are sufficiently uncorrelated. Prior to conducting and building the full models, I had presumed the microtonal melodies experiment (Ch. 4) would distinguish between these two models. In fact, it turned out the data produced by spectral and virtual pitch models were highly correlated ($r(108) = .95$, as shown in Sec. 4.3.2.4). Different experimental stimuli are required to ensure these two models' predictions are less highly correlated.

In Chapter 5, I made some additional hypotheses to account for additional aspects of tonality, such as why the diatonic scale's Ionian tonic is stronger than its Aeolian (natural minor) tonic, and how the tensions and stabilities of pitches may be affected by their chord degree as well as their scale degree. However, I have not provided a formal mathematical model for these hypotheses. In the next section, I briefly outline another hypothesis that may account for these features, and which deals with successions of chords rather than the relationships between a given scale and its chords. It is my intention to create a model that embodies all of these proposed processes.

However, an additional problem is that there is a lack of experimental data encapsulating many important tonal effects. The probe tone data measure certain important aspects of tonal perception and, as shown earlier, I have successfully modelled this data. But they do

1 Most bottom-up models of consonance/dissonance (e.g., Sethares (1993)) can make predictions for microtonal intervals and chords, but they are not designed to model the effects produced by successive pitches and chords.

not capture aspects such as: (a) the effects of differing orderings of any given set of chords; (b) the effects of a wide variety of different chord progressions; (c) the effects of different types of voice-leading; (d) the effects of chord choice, irrespective of their aggregate pitch class content (e.g., the chord progressions Fmaj–Gmaj–Cmaj and Dmin–Emin–Cmaj have very different tonal effects, but both have the same aggregate pitch content). Furthermore, the use of octave (rather than harmonic) complex tones may have unintended consequences.

In the next section, I discuss ways in which some of these limitations can be overcome in future work.

6.3 FUTURE WORK

In order to gain a deeper understand of tonality, there is a vital need to obtain more experimental data; particularly data that provide more detailed information about the effects mentioned above. In earlier work, I conducted a small experiment designed to illuminate these properties (Milne, 2009b,a). Thirty-five participants were asked to rate—on a seven-point scale—the degree of closure produced by the final chord in a variety of three-chord progressions. In order to minimize the effects of consonance/dissonance, all chords were major or minor triads in root position. To maximize ecological validity, realistic sounding timbres were used—the chords were played by a (sampled) string quartet—and conventional rules of voice-leading were followed (i.e., minimizing the voice-leading distance, whilst avoiding parallel octaves and fifths and, as much as possible, hidden octaves and fifths).[2] This experiment provided very useful data, and the participants' responses were highly correlated with each other (mean inter-participant correlation of $r(33) = .49$, Cronbach's $\alpha = .97$) (Milne, 2009a). However, the stimuli were limited in scope. Ignoring voice-leading, and trans-

2 It is widely believed that parallel and hidden fifths are avoided in common practice music in order to preserve the perceived independence of the voices (Huron, 2001). *Hidden* octaves and fifths occur when two voices move in similar motion to an octave or fifth. They are generally considered less objectionable than parallel octaves and fifths, particularly when one or both voices are the alto and tenor (inner) voices.

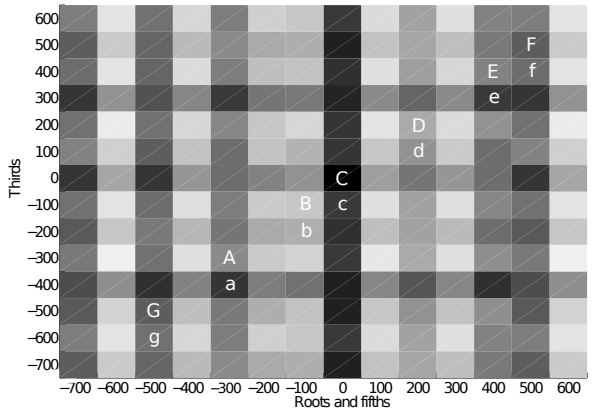
position with respect to the final chord, there are $24 \times 24 \times 2 = 1152$ different such chord progressions; I tested only 72 (6%) of these.

In order to gain more extensive data, I recently conducted a much more comprehensive version of the experiment. I asked each participant (from a total of 120 participants) to rate a random selection of 128 chord progressions from a total of 642. This means each of the 642 progressions were rated, on average, by $120 \times 128/642 \approx 24$ participants. The 642 progressions constitute more than half of all possible such progressions; furthermore, the 510 excluded progressions were those which traverse a large distance in the cycle-of-fifths and can, therefore, be reasonably considered rare in Western music. I have not yet had a chance to fully analyze or model this data. But it provides a hugely powerful resource of information for future research and, notably, to test models of tonality that are designed to capture the effects of chord ordering, voice-leading, and chord degree.

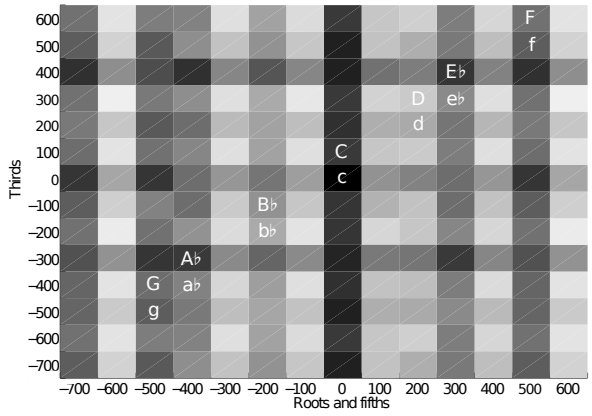
In particular, I intend to use these data to test the hypotheses I made in Chapter 5 (and reiterated above). Namely that (a) given a scale, the strongest sense of harmonic resolution is induced when a bad-fitting tone moves by a small interval (e.g., a semitone) to the root of a best-fitting chord; (b) there is a feedback mechanism whereby the increased salience of a tonic degree affects the instability of its leading tone(s); (c) the tension of pitches is a function of their affinity with their local harmonic context (chord degree) in addition to their scalic context (scale degree). Of course, the first task is to formally embody these principles in a mathematical model so they can be tested against these data.

However, there is an additional novel hypothesis I wish to mention. This hypothesis may also account for the above tonal effects and others, and these data are ideal to test it. It can be exemplified by returning to the figure introduced in Chapter 3. I reproduce this figure, and its minor version, in Figure 6.1.

The hypothesis has two components. Firstly, given two chords played in sequence, we may hear one or both of these chords as perturbations (alterations) of similar chords that have greater spectral pitch



(a) Major reference triad.



(b) Minor reference triad.

Figure 6.1: Spectral pitch similarities of a Cmaj or Cmin reference triad and all possible 12-TET triads that contain a perfect fifth. Spectral pitch similarity is calculated with the previously optimized smoothing of 10.3 cents and roll-off of 0.42.

similarity. For instance (and as shown in Fig. 6.1a), the chord pair Cmaj–Dmaj has lower spectral pitch similarity than the similar chord pair Cmaj–Dmin. The hypothesis implies that when the (lower affinity) Cmaj–Dmaj is played, it may be heard as a perturbation of the (higher affinity) Cmaj–Dmin. More specifically, the played pitch F♯ may be heard as an alteration of the pitch F.

Secondly, when a pitch is heard as perturbed in this way, it is heard to resolve when it continues—in the same direction as its perturbation—to the next available pitch with high affinity. For example, in the chord progression Cmaj–Dmaj–Gmaj, the tone F♯ is resolved by moving to G (i.e., in the same direction as its (upwards) perturbation from F). In

this way, the F♯ acts as a low affinity passing tone between an F that is implied (due to its higher affinity) and the played G, which follows.

This may account for the perceived instability or activity of the V in the IV–V progression (of which Cmaj–Dmaj is an example), and for its perceived need to resolve to I (the Gmaj chord). Numerous similar examples can be found; for example, the spectral pitch similarity of Cmaj and F♯maj is lower than that of Cmaj and F♯min, which suggests that, when the former is played, the tone A♯ may be heard as an active alteration of the higher affinity tone A and, hence, seeks resolution to B in the cadence Cmaj–F♯maj–Bmin (a ♭II–V–i cadence). Further examples are provided in Milne et al. (2011b).

By combining all these additional hypotheses, it may be possible to model—from the bottom up—the feelings of tension, activity, resolution, and so forth, that are amongst the most characteristic features of harmonic tonality. A model to explore this bottom-up account of tonal functionality is in progress, and I plan to report on its effectiveness in future publications.

The work described in this dissertation can also be extended in other ways. To date, I have not used any of the relative or higher-dimensional tensors in formal models of music cognition. In Sections 3.6.2 and 3.6.3, I showed how such expectation tensors provide effective methods for calculating the similarity between any arbitrary tuning of a scale and just intonation (or other privileged) referents, and how they can generalize many of the methods used in musical set theory. It will be interesting to see how such tensors may become useful in models of tonal perception. For instance, Kuusi (2001) collected ratings from 58 participants for the perceived 'closeness' of chords; these data could be modelled by a linear combination of the distances between a variety of tensors of differing orders.

Another research opportunity deriving from the tensors is to develop methods for creating *salience* rather than expectation tensors. In Section 3.4, I showed how expectation tensors sum the elements $x[i, j]$ in the pitch response matrix $\mathbf{X}$ to give the expected numbers of tones

(or ordered tone pairs, tone triples, etc.) heard at any given pitch (or dyad, triad, etc.). An alternative strategy is to derive the probability of hearing any given pitch, dyad, triad, and so forth (regardless of how many tones, or tuples of tones, we may hear playing it). The resulting tensors will have elements with values in the interval $[0, 1]$. It is quite straightforward to naively calculate the elements of such a tensor. For example, the elements of an absolute monad salience tensor are given by $x_{\mathrm{s}}[j] = 1 - \prod_{i=1}^{I} 1 - x[i, j]$ (compare this with the absolute monad expectation tensor, which is $x_{\mathrm{e}}[j] = \sum_{i=1}^{I} x[i, j]$). However, the method of inclusion-exclusion of tensor subspaces, which greatly reduces the computational complexity of the higher-order tensors, cannot be directly replicated for the salience tensors. Despite this, it should be possible to find a related method of inclusion-exclusion that will provide similar reductions in computational complexity.

A further research opportunity opened up by the expectation tensors is to build in the option for an additional form of invariance—*scale invariance*. Currently, the tensors can embed pitches or pitch classes (only the latter have invariance with respect to the octave), and these embeddings can be represented in an absolute or a relative form (only the latter has invariance with respect to transposition). Scale invariance may also have useful cognitive applications—under a relative dyad embedding that is also scale invariant, the representations of, for example, pitch sets {C4, D4, E♭4} and {C4, E4, F♯4} would be identical—the latter is essentially a scaled (stretched) version of the former (whole tones become major thirds, and semitones become whole tones). With a means to indicate the temporal ordering of the embedded intervals, they would become contour invariant—melodies with the same contour 'shapes' but differing contour 'depths' would be invariant under such an embeddings. Melodies with similar contours would have similar embeddings under a standard metric. I have some preliminary ideas about how such embeddings may be constructed.

A further research question that could be tackled is to devise experimental stimuli that can more clearly distinguish between the effects

of virtual and spectral pitch similarity. One way to achieve this would be to create synthetic timbres whose modelled virtual pitches are distinctly different to their spectral pitches. Participants would then judge the affinities of differently sized intervals that match either the spectral or the virtual pitches.

And, on the distant horizon is the grand aim of constructing a complete model of tonality comprising submodels for each of the four categories of mental processes illustrated in Figure 2.1: intrinsic, extrinsic, nature, and nurture.

6.4 IMPLICATIONS

Understanding the extent to which the cultural artefacts of humankind are due to innate processes of human perception and cognition, and how much they are down to learning is a key question. In this dissertation, I have posited that innate processes play an important role in the shaping of tonal music.

So, does this mean there is only one true form of music, one *ideal* to which artists may aspire, and against which existing pieces must be judged? No. Rather, it just suggests likely iconic significations for certain musical events. For example, we can argue there is a likely association between the 'discomfort' induced by dissonance, or poor affinity, and negatively valenced feelings. However, existing symbols and the contexts within which they operate are often deliberately subverted by the artist—who seeks to make the familiar, unfamiliar.[3] The cultural context is also important—the zeitgeist may consider the 'discomfort' of disaffinity or dissonance as a positive aesthetic, and a sophisticated audience may feel that excessive consonance and affinity is too simplistic, too unlike 'reality', to communicate in an authentic or meaningful manner.

3 This process is known as *defamiliarization*; Shklovsky (1965, p. 11–12) writes, 'the technique of art is to make objects "unfamiliar," to make forms difficult, to increase the difficulty and length of perception because the process of perception is an aesthetic end in itself and must be prolonged', (as quoted in Thompson (1988)).

In the end, it is the artist—his or her aesthetic sensibilities—and the culture within which he or she operates that dictates how the loose emotional connotations of musical events may become transformed and transfigured into the remarkable art form that is music: 'All good music resembles something. Good music stirs by its mysterious resemblance to the objects and feelings which motivated it' Cocteau (1918, p. 51).[4]

4 Toute bonne musique est *ressemblante*. La bonne musique émeut par cette ressemblance mystérieuse avec les objets et les sentiments qui l'ont motivée.

REFERENCES

Agmon, E. (1995). Functional harmony revisited: A prototype-theoretic approach. *Music Theory Spectrum*, 17(2):196–214. (Cited on page 42.)

Balzano, G. J. (1980). The group-theoretic description of 12-fold and microtonal pitch systems. *Computer Music Journal*, 4(4):66–84. (Cited on pages 43 and 44.)

Barbour, J. M. (1951). *Tuning and Temperament: A Historical Survey*. Michigan State College Press, East Lansing, Michigan. (Cited on page 184.)

Berger, K. (1987). *Musica Ficta: Theories of Accidental Inflections in Vocal Polyphony from Marchetto da Padova to Gioseffo Zarlino*. Cambridge University Press. (Cited on page 175.)

Berlyne, D. E. (1970). Novelty, complexity and hedonic value. *Perception and Psychophysics*, 8(5):279–286. (Cited on page 29.)

Bernstein, J. G. and Oxenham, A. J. (2003). Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? *The Journal of the Acoustical Society of America*, 113(6):3323–3334. (Cited on page 130.)

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., and D'Adamo, D. A. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, 29(1):159–171. (Cited on page 22.)

Bohlen, H. (1978). 13 tonstufen in der duodezime. *Acustica*, 39(2):76–86. (Cited on page 91.)

Bregman, A. S. (1990). *Auditory Scene Analysis*. MIT Press, Cambridge, MA, USA. (Cited on page 98.)

Brown, H. (1988). The interplay of set content and temporal context in a functional theory of tonality perception. *Music Perception*, 5(3):219–250. (Cited on page 46.)

Brown, J. C. (1992). Musical fundamental frequency tracking using a pattern recognition method. *The Journal of the Acoustical Society of America*, 92(3):1394–1402. (Cited on page 107.)

Buchler, M. (2001). Relative saturation of intervals and set classes: A new approach to complementation and pcset similarity. *Journal of Music Theory*, 45(2):263–343. (Cited on pages 77 and 91.)

Bucht, S. and Huovinen, E. (2004). Perceived consonance of harmonic intervals in 19-tone equal temperament. In Parncutt, R., Kessler, A., and Zimmer, F., editors, *Proceedings of the Conference on Interdisciplinary Musicology*, Graz, Austria. (Cited on page 195.)

Budrys, R. and Ambrazevicius, R. (2008). 'Tonal' vs 'atonal': Perception of tonal hierarchies. In E. Cambouropoulos, E. C., Parncutt, R., Solomos, M., Stefanou, D., and Tsougras, C., editors, *Proceedings of the 4th Conference on Interdisciplinary Musicology*, pages 36–37, Aristotle University, Thessaloniki, Greece. (Cited on page 150.)

Busemeyer, J. R. and Diederich, A. (2010). *Cognitive Modeling*. SAGE, Los Angeles, USA. (Cited on page 21.)

Butler, D. (1989). Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry. *Music Perception*, 6(3):219–242. (Cited on pages 151, 157, and 171.)

Caplin, W. (1983). Tonal function and metrical accent: A historical perspective. *Music Theory Spectrum*, 5(1):1–14. (Cited on page 42.)

Carey, N. (2007). Coherence and sameness in well-formed and pairwise well-formed scales. *Journal of Mathematics and Music*, 1(2):79–98. (Cited on page 86.)

Carey, N. and Clampitt, D. (1989). Aspects of well-formed scales. *Music Theory Spectrum*, 11(2):187–206. (Cited on pages 86 and 177.)

Castrén, M. (1994). *Recrel: A Similarity Measure for Set-Classes*. Sibelius Academy, Helsinki. (Cited on pages 51, 77, and 91.)

Chalmers, J. (1990). *Divisions of the Tetrachord*. Frog Peak Music. (Cited on pages 57 and 145.)

Chandler, D. (2002). *Semiotics: The Basics*. Routledge, London. (Cited on page 10.)

Chen, S., Ma, B., and Zhang, K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, 410(24–25):2365–2376. (Cited on page 55.)

Chew, E. (2000). *Towards a mathematical model of tonality*. PhD dissertation, Massachusetts Institute of Technology. (Cited on page 44.)

Chew, E. (2005). Real-time pitch spelling using the spiral array. *Computer Music Journal*, 29(2):61–76. (Cited on page 44.)

Chew, E. (2006). Slicing it all ways: mathematical models for tonal induction, approximation, and segmentation using the spiral array. *INFORMS Journal on Computing*, 18(3):305. (Cited on page 44.)

Cocteau, J. (1918). *Le coq et l'arlequin: notes autour de la musique*. Éditions de la Sirène, Paris. (Cited on page 203.)

Cuddy, L. L. and Thompson, W. F. (1992). Asymmetry of perceived key movement in chorale sequences: Converging evidence from a probe-tone analysis. *Psychological Research*, 54(2):51–59. (Cited on page 46.)

Dahlhaus, C. (1980). Tonality. In Sadie, S., editor, *The New Grove Dictionary of Music and Musicians*, volume 19, pages 51–55. Macmillan, London, 1st edition. (Cited on page 7.)

Dahlhaus, C. (1990). *Studies on the Origin of Harmonic Tonality*. Princeton University Press, Oxford. (Cited on pages 7 and 35.)

Dennett, D. C. (1994). Cognitive science as reverse engineering: Several meanings of "top-down" and "bottom-up". In Prawitz, D., Skyrms, B., and Westerståhl, D., editors, *Logic, Methodology, and Philosophy of Science IX*, pages 679–689, Amsterdam. Elsevier Science. (Cited on page 22.)

Deutsch, D. (1997). *The Fabric of Reality: Towards a Theory of Everything*. Penguin Books, London. (Cited on pages 19 and 20.)

Dretske, F. (2006). Perception without awareness. In Gendler, T. S. and Hawthorne, J., editors, *Perceptual Experience*. Oxford University Press, Oxford, UK. (Cited on page 53.)

Erlich, P. (1998). Tuning, tonality, and twenty-two-tone temperament. *Xenharmonikôn*, 17. (Cited on page 186.)

Erlich, P. (2006). A middle path between just intonation and the equal temperaments, part 1. *Xenharmonikôn*, 18:159–199. (Cited on pages 86, 90, 174, and 185.)

Euler, L. (1739). *Tentamen novae theoriae musicae ex certissismis harmoniae principiis dilucide expositae*. Saint Petersburg Academy, Saint Petersburg. (Cited on page 43.)

Ferguson, S., Schubert, E., and Dean, R. T. (2011). Continuous subjective loudness responses to reversals and inversions of a sound recording of an orchestral excerpt. *Musicae Scientiae*, 15(3):387–401. (Cited on page 22.)

Ferrer, R. (2007). *The role of timbre in the memorization of microtonal intervals*. Master's, University of Jyväskylä. (Cited on page 195.)

Fétis, F. J. (1844). *Traité complet de la théorie et de la pratique de l'harmonie*. Schlesinger, Paris. (Cited on page 9.)

Forte, A. (1973). *The Structure of Atonal Music*. Yale University Press. (Cited on pages 72, 77, 91, 92, and 93.)

Gell, A. (1995). The language of the forest: Landscape and phonological iconism in Umeda. In Hirsch, E. and O'Hanlon, M., editors, *The Anthropology of Landscape*, pages 232–254. Clarendon Press, Oxford. (Cited on page 24.)

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54:1496–1516. (Cited on page 62.)

Green, B. and Butler, D. (2002). From acoustics to *Tonpsychologie*. In Christensen, T., editor, *The Cambridge History of Western Music Theory*, pages 246–271. Cambridge University Press, Cambridge. (Cited on page 98.)

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York. (Cited on page 62.)

Harrison, D. (1994). *Harmonic Function in Chromatic Music: A Renewed Dualist Theory and an Account of its Precedents*. University of Chicago Press, Chicago and London. (Cited on pages 42 and 183.)

Helmholtz, H. L. F. (1877). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Dover, New York. (Cited on pages 98 and 161.)

Heyduk, R. G. (1975). Rated preference for music composition as it relates to complexity and exposure frequency. *Perception and Psychophysics*, 17:84–91. (Cited on page 29.)

Honing, H. (2006). Computational modeling of music cognition: A case study on model selection. *Music Perception*, 23(5):365–376. (Cited on page 17.)

Hostinský, O. (1879). *Die Lehre von den musikalischen Kl angen: Ein Beitrag zur aesthetischen Beg urndung der Harmonie-Lehre*. H. Dominicus, Prague. (Cited on page 43.)

Huron, D. (1991). Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception*, 9(2):135–154. (Cited on pages 13 and 98.)

Huron, D. (1994). Interval-class content in equally tempered pitch-class sets: Common scales exhibit optimum tonal consonance. *Music Perception*, 11(3):289–305. (Cited on pages 13, 145, 162, and 163.)

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, 19(1):1–64. (Cited on pages 13, 110, 118, and 197.)

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, MA, USA. (Cited on pages 2, 10, 11, 12, 13, and 46.)

Hutchinson, W. and Knopoff, L. (1978). The acoustic component of western consonance. *Interface*, 7:1–29. (Cited on pages 161 and 163.)

Hyer, B. (2001). Tonality. In Sadie, S. and Tyrrell, J., editors, *The New Grove Dictionary of Music and Musicians*, volume 25, pages 583–594. Macmillan, London, 2nd edition. (Cited on page 7.)

Hyer, B. (2002). Tonality. In Christensen, T., editor, *The Cambridge History of Western Music Theory*, pages 726–752. Cambridge University Press, Cambridge. (Cited on pages 7 and 10.)

Jairazbhoy, N. A. (1995). *The Rags of North Indian Music: Their Structure and Evolution*. Popular Prakashan, 1st revised edition. (Cited on page 9.)

Jerket, J. (2004). Music articulation in the organ. In *Proceedings of Joint Baltic-Nordic Acoustics Meeting*, Mariehamn, Åland, Finland. (Cited on page 121.)

Juhász, Z. (2012). A mathematical study of note association paradigms in different folk music cultures. *Journal of Mathematics and Music*, 6(3):169–185. (Cited on page 105.)

Juslin, P. N. and Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238. (Cited on pages 10 and 27.)

Juslin, P. N. and Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioural and Brain Sciences*, 31:559–621. (Cited on pages 24, 26, 27, 29, and 194.)

Kameoka, A. and Kuriyagawa, M. (1969). Consonance theory parts 1 and 2. *The Journal of the Acoustical Society of America*, 45(6):1451–1469. (Cited on pages 109, 161, and 163.)

Kelley, R. T. (2004). Reconciling tonal conflicts: Mod-7 transformations in chromatic music. In *Paper delivered at the second annual meeting of the Music Theory Society of the Mid Atlantic in Philadelphia, Pennsylvania*. (Cited on page 42.)

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53(11):3735–3745. (Cited on page 136.)

Korte, B. and Vygen, J. (2007). *Combinatorial Optimization: Theory and Algorithms*. Springer, Berlin, 4th edition. (Cited on page 56.)

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford. (Cited on pages 2, 37, 38, 40, 42, 44, 45, 46, 81, 98, 149, 150, 151, 154, 158, 159, and 161.)

Krumhansl, C. L. and Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4):334–368. (Cited on pages xxiv, 6, 147, 148, 149, 150, 153, and 158.)

Krumhansl, C. L. and Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, 5:579–594. (Cited on page 150.)

Kuusi, T. (2001). *Set-class and chord: Examining connection between theoretical resemblance and perceived closeness*. PhD thesis, Sibelius Academy, Helsinki, Finland. (Cited on pages 77, 91, 92, and 200.)

Lansberg, M. E. (1980). The icon in semiotic theory. *Current Anthropology*, 21(1):93–95. (Cited on page 24.)

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, 17(4):481–509. (Cited on pages 151, 160, and 171.)

Lerdahl, F. (1988). Tonal pitch space. *Music Perception*, 5(3):315–350. (Cited on page 156.)

Lerdahl, F. (2001). *Tonal Pitch Space*. Oxford University Press, Oxford. (Cited on pages 2, 42, 98, and 156.)

Lewandowski, S. and Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. SAGE, Los Angeles, USA. (Cited on page 20.)

Lewin, D. (1959). Re: Intervallic relations between two collections of notes. *Journal of Music Theory*, 3(2):298–301. (Cited on page 71.)

Lewin, D. (1987). *Generalized Musical Intervals and Transformations*. Yale University Press, New Haven, CT. (Cited on page 77.)

Lewin, D. (2001). Special cases of the interval function between pitch-class sets X and Y. *Journal of Music Theory*, 45:1–29. (Cited on page 71.)

Lindley, M. (2013). Pythaorean intonation. In *Grove Music Online. Oxford Music Online*. Oxford University Press. (Cited on page 176.)

Longuet-Higgins, H. C. (1962). Two letters to a musical friend. *The Music Review*, 23:244–248. (Cited on page 43.)

Lowinsky, E. E. (1961). *Tonality and Atonality in Sixteenth-Century Music*. University of California Press, Los Angeles. (Cited on page 8.)

MacCallum, R. M., Mauch, M., Burt, A., and Leroi, A. M. (2012). Evolution of music by public choice. *Proceedings of the National Academy of Sciences of the United States of America*. (Cited on page 34.)

Macpherson, S. (1920). *Melody and Harmony: A Treatise for the Teacher and the Student*. Joseph Williams, London. (Cited on pages 46, 177, 180, and 181.)

Malmberg, C. F. (1918). The perception of consonance and dissonance. *Psychological Monographs*, 25(2):93–133. (Cited on pages 161 and 163.)

Mashinter, K. (2006). Calculating sensory dissonance: Some discrepancies arising from the models of Kameoka & Kuriyagawa, and Hutchinson & Knopoff. *Empirical Musicology Review*, 1(2):65–84. (Cited on page 109.)

Mathews, M. V., Roberts, L. A., and Pierce, J. R. (1984). Four new scales based on nonsuccessive-integer-ratio chords. *Journal of the Acoustical Society of America*, 75(S1):S10. (Cited on page 91.)

Mathieu, W. A. (1997). *Harmonic Experience*. Inner Traditions, Rochester. (Cited on page 42.)

McDermott, J. H., Lehr, A. J., and Oxenham, A. J. (2010). Individual differences reveal the basis of consonance. *Current Biology*, 20(11):1035–1041. (Cited on page 98.)

Meyer, L. B. (1956). *Emotion and Meaning in Music*. University of Chicago Press, Chicago. (Cited on pages 2 and 10.)

Milne, A. J. (2009a). A psychoacoustic model of harmonic cadences. Master's thesis, University of Jyväskylä, Finland. (Cited on pages 41, 49, 51, 83, 85, 105, and 197.)

Milne, A. J. (2009b). A psychoacoustic model of harmonic cadences: A preliminary report. In Louhivuori, J., Eerola, T., Saarikallio, S., Himberg, T., and Eerola, P., editors, *ESCOM 2009 Proceedings*, pages 328–337, Jyväskylä, Finland. (Cited on page 197.)

Milne, A. J. (2010). Tonal music theory: A psychoacoustic explanation? In Demorest, S. M., Morrison, S. J., and Campbell, P. S.,

editors, *Proceedings of the 11th International Conference on Music Perception and Cognition*, pages 597–600, University of Washington, Seattle, USA. (Cited on page 51.)

Milne, A. J., Carlé, M., Sethares, W. A., Noll, T., and Holland, S. (2011a). Scratching the scale labyrinth. In Agon, C., Amiot, E., Andreatta, M., Assayag, G., Bresson, J., and Mandereau, J., editors, *Mathematics and Computation in Music – MCM 2011*, volume 6726 of *LNAI*, pages 180–195, Berlin. Springer-Verlag. (Cited on page 174.)

Milne, A. J. and Prechtl, A. (2008). New tonalities with the Thummer and The Viking. In Crossan, A. and Kaaresoja, T., editors, *Proceedings of the 3rd International Haptic and Auditory Interaction Design Workshop*, volume 2, pages 20–22, Jyväskylä, Finland. (Cited on page 116.)

Milne, A. J., Sethares, W. A., Laney, R., and Sharp, D. B. (2011b). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music*, 5(1):1–20. (Cited on pages 56 and 200.)

Milne, A. J., Sethares, W. A., and Plamondon, J. (2007). Isomorphic controllers and Dynamic Tuning: Invariant fingering over a tuning continuum. *Computer Music Journal*, 31(4):15–32. (Cited on page 86.)

Milne, A. J., Sethares, W. A., and Plamondon, J. (2008). Tuning continua and keyboard layouts. *Journal of Mathematics and Music*, 2(1):1–19. (Cited on pages 86, 89, and 117.)

Moore, B. C. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America*, 54:610–619. (Cited on page 243.)

Moore, B. C., Glasberg, B. R., and Shailer, M. J. (1984). Frequency and intensity difference limens for harmonics within complex tones. *The Journal of the Acoustical Society of America*, 75(2):500–561. (Cited on pages 221 and 243.)

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259. (Cited on page 33.)

North, A. C. and Hargreaves, D. J. (1995). Subjective complexity, familiarity, and liking for popular music. *Psychomusicology*, 14:77–93. (Cited on pages 14, 29, and 30.)

Parncutt, R. (1988). Revision of Terhardt's psychoacoustical model of the root(s) of a musical chord. *Music Perception*, 6(1):65–94. (Cited on pages 45, 106, 107, 163, 164, and 165.)

Parncutt, R. (1989). *Harmony: A Psychoacoustical Approach*. Springer-Verlag. (Cited on pages 51, 80, 83, 99, and 159.)

Parncutt, R. (1992). How "middle" is middle C? Terhardt's virtual pitch weight and the distribution of pitches in music. Unpublished. (Cited on page 118.)

Parncutt, R. (1994). Template-matching models of musical pitch and rhythm perception. *Journal of New Music Research*, 23:145–167. (Cited on page 166.)

Parncutt, R. (2011). The tonic as triad: Key profiles as pitch salience profiles of tonic triads. *Music Perception*, 28(4):333–365. (Cited on pages 150, 151, 152, 158, 163, 166, 171, 174, and 176.)

Parncutt, R. and Hair, G. (2011). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, 5(2):119–166. (Cited on page 98.)

Parncutt, R. and Prem, D. (2008). The relative prevalence of medieval modes and the origin of the leading tone. Poster presented at International Conference of Music Perception and Cognition (ICMPC10), Sapporo, Japan, 25–29 August 2008. (Cited on pages 166 and 174.)

Pearce, M. T. and Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405. (Cited on pages 2, 22, 37, 38, 40, 98, and 108.)

Peterson, L. R. and Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58:193–198. (Cited on pages 24 and 158.)

Petitmengin, C. and Bitbol, M. (2009). The validity of first-person descriptions as authenticity and coherence. *Journal of Consciousness Studies*, 16(10–12):363–404. (Cited on pages 32 and 33.)

Piston, W. and Devoto, M. (1987). *Harmony*. Norton, New York, 5th edition. (Cited on pages 46, 177, 178, 179, and 181.)

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3):472–491. (Cited on page 17.)

Plomp, R. and Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4):548–560. (Cited on pages xx, 109, 113, and 144.)

Pratt, G. (1996). *The Dynamics of Harmony: Principles and Practice*. Oxford University Press, Oxford, UK. (Cited on pages 46 and 179.)

Prechtl, A., Milne, A. J., Holland, S., Laney, R., and Sharp, D. B. (2012). A MIDI sequencer that widens access to the compositional possibilities of novel tunings. *Computer Music Journal*, 36(1):42–54. (Cited on page 177.)

Quinn, I. (2005). Harmonic function without primary triads. (Cited on page 42.)

Ramachandran, V. S. and Hubbard, E. M. (2001). Synaesthesia—a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34. (Cited on page 24.)

Rameau, J.-P. (1726). *Nouveau système de musique théorique*. Paris. (Cited on page 42.)

Riemann, H. (1914–1915). Ideen zu einer 'lehre von den tonvorstellungen'. *Jahrbuch der Bibliothek Peters*, 21–22(1–26). (Cited on page 43.)

Roederer, J. G. (2008). *The Physics and Psychophysics of Music*. Springer, 4th edition. (Cited on pages 45 and 52.)

Rogers, D. W. (1999). A geometric approach to pcset similarity. *Perspectives of New Music*, 37:77–90. (Cited on page 81.)

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121. (Cited on page 56.)

Schenker, H. (1954). *Harmony*. University of Chicago Press, Chicago. (Cited on page 42.)

Schenker, H. (1987). *Counterpoint*. Schirmer, New York. (Cited on page 42.)

Schoenberg, A. (1969). *Structural Functions of Harmony*. Faber and Faber, London, 2nd edition. (Cited on page 182.)

Schoenberg, A. (1978). *Theory of Harmony*. Faber and Faber, London, 3rd revised edition. (Cited on page 9.)

Scott, D. and Isaacson, E. J. (1998). The interval angle: A similarity measure for pitch-class sets. *Perspectives of New Music*, 36:107–42. (Cited on page 81.)

Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, 94(3):1218–1228. (Cited on pages 109 and 196.)

Sethares, W. A. (2005). *Tuning, Timbre, Spectrum, Scale*. Springer Verlag, London, 2nd edition. (Cited on pages xviii, xx, 102, 109, 113, 114, and 144.)

Sethares, W. A., Milne, A. J., Tiedje, S., Prechtl, A., and Plamondon, J. (2009). Spectral tools for Dynamic Tonality and audio morphing.

Computer Music Journal, 33(2):71–84. (Cited on pages 59, 116, 117, and 143.)

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7(221–264). (Cited on page 18.)

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323. (Cited on page 55.)

Shklovsky, V. (1965). Art as technique. In Lemon, L. T. and Reis, M. J., editors, *Russian Formalist Criticism: Four Essays*, pages 3–24. University of Nebraska Press. (Cited on page 202.)

Sloboda, J. A. and Juslin, P. N. (2001). *Music and Emotion*, chapter Psychological perspectives on music and emotion, pages 71–104. Oxford University Press, Oxford. (Cited on pages 22, 24, and 194.)

Smith, A. B. (1997). A "cumulative" method of quantifying tonal consonance in musical key contexts. *Music Perception*, 15(2):175–188. (Cited on page 162.)

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1):89–96. (Cited on page 30.)

Steinke, W. R., Cuddy, L. L., and Holden, R. R. (1993). Perception of musical tonality as assessed by the probe-tone method. *Canadian Acoustics*, 21:85–86. (Cited on page 150.)

Steinke, W. R., Cuddy, L. L., and Holden, R. R. (1997/1998). Dissociation of musical tonality and pitch memory from nonmusical cognitive abilities. *Canadian Journal of Experimental Psychology*, 51:316–334. (Cited on page 150.)

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society*, 39(1):44–47. (Cited on page 18.)

Stumpf, C. (1890). *Tonpsychologie*, volume 2. Leipzig. (Cited on page 98.)

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66. (Cited on page 126.)

Sun, R. (2001). *Duality of the Mind: A Bottom-Up Approach Toward Cognition*. Psychology Press. (Cited on page 22.)

Tenney, J. (1988). *A History of 'Consonance' and 'Dissonance'*. Excelsior Music, New York. (Cited on pages 14, 15, 43, 99, and 145.)

Terhardt, E. (1984). The concept of musical consonance: A link between music and psychoacoustics. *Music Perception*, 1(3):276–295. (Cited on page 98.)

Terhardt, E., Stoll, G., and Seewann, M. (1982). Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions. *The Journal of the Acoustical Society of America*, 71(3):671–678. (Cited on pages 45, 51, 83, and 106.)

Thompson, K. (1988). *Breaking the Glass Armor: Neoformalist Film Analysis*. Princeton University Press. (Cited on page 202.)

Thompson, W. F. (1986). *Judgements of key change in Bach chorale excerpts: An investigation of the sensitivity to keys, chords, and voicing*. PhD dissertation, Queen's University at Kingston. (Cited on page 150.)

Toiviainen, P. and Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, 32(6):741–766. (Cited on pages 46, 150, and 171.)

Trout, J. D. (2007). The psychology of scientific explanation. *Philosophy Compass*, 2(3):564–591. (Cited on page 20.)

Tymoczko, D. (2006). Supporting online material for the geomtery of musical chords. *Science*. (Cited on pages 51 and 57.)

Vitz, P. C. (1966). Affect as a function of stimulus variation. *Journal of Experimental Psychology*, 71:74–79. (Cited on page 29.)

von Oettingen, A. (1866). *Harmoniesystem in dualer Entwicklung*. Dorpat. (Cited on page 43.)

Vos, J. (1982). The perception of pure and mistuned musical fifths and major thirds: Thresholds for discrimination and beats. *Perception and Psychophysics*, 32(4):297–313. (Cited on page 195.)

Vos, J. (1984). Spectral effects in the perception of pure and tempered intervals: Discrimination and beats. *Perception and Psychophysics*, 35(2):173–185. (Cited on page 195.)

Vos, J. (1986). Purity ratings of tempered fifths and major thirds. *Music Perception*, 3(3):221–257. (Cited on page 195.)

Vos, J. (1988). Subjective acceptability of various regular twelve-tone tuning systems in two-part musical fragments. *Journal of the Acoustical Society of America*, 83(6):2383–2392. (Cited on page 195.)

Vos, J. and van Vianen, B. G. (1985a). The effect of fundamental frequency on the discriminablity between pure and tempered fifths and major thirds. *Perception and Psychophysics*, 37(6):507–514. (Cited on page 195.)

Vos, J. and van Vianen, B. G. (1985b). Thresholds for discrimination between pure and tempered intervals: The relevance of nearly coinciding harmonics. *Journal of the Acoustical Society of America*, 77(1):176–187. (Cited on page 195.)

Vos, P. G. and Troost, J. M. (1989). Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception*, 6(4):383–396. (Cited on page 118.)

Wescott, R. W. (1971). Linguistic iconism. *Language*, 47(2):416–428. (Cited on page 24.)

Widdess, R. (1981). Aspects of form in North Indian alap and dhrupad. In Widdess, D. R., editor, *Music and Tradition: Essays on Asian & Other Musics Presented to Lauren Picken*, pages 143–182. Cambridge University Press, Cambridge, UK. (Cited on page 9.)

Wilding-White, R. (1961). Tonality and scale theory. *Journal of Music Theory*, 5(2):275–286. (Cited on page 163.)

Wilson, E. (1975). Letter to Chalmers pertaining to moments-of-symmetry/Tanabe cycle. Unpublished. (Cited on pages 86 and 177.)

Woolhouse, M. (2007). *Interval cycles and the cognition of pitch attraction in Western tonal-harmonic music*. PhD dissertation, Cambridge University, Department of Music, Cambridge. (Cited on page 46.)

Woolhouse, M. (2009). Modelling tonal attraction between adjacent musical elements. *Journal of New Music Research*, 38(4):357–379. (Cited on pages 22, 37, and 144.)

Woolhouse, M. and Cross, I. (2010). Using interval cycles to model Krumhansl's tonal hierarchies. *Music Theory Spectrum*, 32(1):60–78. (Cited on page 158.)

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44(1):41–61. (Cited on page 17.)

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: Facts and models*. Springer, Berlin. (Cited on pages 51 and 83.)

A

SMOOTHING WIDTH AND THE DIFFERENCE LIMEN

The *frequency difference limen*, also known as the *just noticeable frequency difference*, is determined in a two-alternative forced-choice (2-AFC) experiment. In this type of experiment, participants are presented with numerous pairs of successively played tones. The tones in each pair either have the same or differing frequencies, and the participant is tasked with categorizing them accordingly.

The frequency difference limen is normally defined as the frequency difference at which the correct response rate indicates a d' (also known as d prime) of approximately one (there are alternative definitions). The value of d' is defined as the distance, in standard deviations, between the mean of the responses to the signal-plus-noise stimuli and the mean of the responses to the noise-alone stimuli (for the above test, a signal-plus-noise stimulus corresponds to two different frequencies; a noise-alone stimulus to two identical frequencies). This implies the internal response to a tone of pitch j can be modelled by a Gaussian centred at j, with a standard deviation σ equivalent to the frequency DL at j.

Experimentally obtained data (e.g., Moore et al. (1984)) typically give a frequency DL, for tones with harmonic partials, that is equivalent (over a broad range of musically useful frequencies) to a pitch DL of approximately 3 cents. Such results are obtained in laboratory conditions with simple stimuli and minimal time gaps between tones (hence comparisons are conducted from auditory sensory (echoic) memory, or short-term memory): in real music, tones and chords are presented as part of a complex and distracting stream of musical information, and there may be long gaps between the presentations of the tone collections (in which time memory of the first pitch may have degraded, so judgements are less precise). For these reasons, it may be appropriate to

treat 3 cents as a minimum standard deviation; larger values may provide more effective results in some models. Indeed, σ can be treated as a free parameter.

B

TENSORS, TENSOR OPERATIONS, AND THEIR NOTATION

A *tensor* is a generalization of a vector or matrix into higher *orders*. An order-0 tensor is a scalar, an order-1 tensor is a vector, an order-2 tensor is a matrix, an order-3 tensor may be thought of as a 3-dimensional array of numbers, and so forth. The *size* of a tensor of order-R denoted $\mathbf{X} \in \mathbb{R}^{I \times J \times \cdots \times M}$ may be shown as $\overbrace{I \times J \times \cdots \times M}^{R}$, which means the first *mode* is of dimension I (it contains I entries); the second mode is of dimension J, and so forth. It is often convenient to specify the order of a tensor by its subscript so that $\mathbf{X}_{J^3}$ represents an order-3 tensor in $\mathbb{R}^{J^3}$ (which is $\mathbb{R}^{J \times J \times J}$). A lowercase italic letter such as $x_{i,j,k}$ or $x[i,j,k]$ denotes a specific scalar element in a tensor, with the subscripts or brackets specifying its location in the tensor. A specific permutation of a tensor's modes is indicated with a subscript in angle brackets, so if $\mathbf{X}$ is a tensor of size $I \times J \times K \times L$, the tensor $\mathbf{X}_{\langle 3,1,4,2 \rangle}$ has size $J \times L \times I \times K$. In other words, $\mathbf{X}_{\langle m_1, m_2, \ldots, m_R \rangle}$ means that the rth mode of $\mathbf{X}$ is moved to the m_rth mode (this is equivalent to the notation used in Mathematica's `Transpose` function; it is not the same as that used in MATLAB's `permute` function).

The symbol $\circ$ denotes the *Hadamard (entrywise) product* of two tensors. If $\mathbf{C} = \mathbf{A} \circ \mathbf{B}$, then $c[i,j,\ldots] = a[i,j,\ldots]\, b[i,j,\ldots]$ ($\mathbf{A}$ and $\mathbf{B}$ must be of the same size). For example,

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \circ \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} = \begin{pmatrix} 1{\cdot}5 & 3{\cdot}7 \\ 2{\cdot}6 & 4{\cdot}8 \end{pmatrix} = \begin{pmatrix} 5 & 21 \\ 12 & 32 \end{pmatrix}. \tag{B.1}$$

The *outer (tensor) product* $\otimes$ of a tensor $\mathbf{A}$ of size $I \times J$ and a tensor $\mathbf{B}$ of size $L \times M$ produces a tensor of size $I \times J \times L \times M$ con-

taining all possible products of their elements. If $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$, then $c[i, j, \ldots, \ell, m, \ldots] = a[i, j, \ldots]\, b[\ell, m, \ldots]$. For example,

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \otimes \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} = \left(\begin{array}{cc|cc} 1{\cdot}5 & 1{\cdot}7 & 3{\cdot}5 & 3{\cdot}7 \\ 1{\cdot}6 & 1{\cdot}8 & 3{\cdot}6 & 3{\cdot}8 \\ \hline 2{\cdot}5 & 2{\cdot}7 & 4{\cdot}5 & 4{\cdot}7 \\ 2{\cdot}6 & 2{\cdot}8 & 4{\cdot}6 & 4{\cdot}8 \end{array}\right) = \left(\begin{array}{cc|cc} 5 & 7 & 15 & 21 \\ 6 & 8 & 18 & 24 \\ \hline 10 & 14 & 20 & 28 \\ 12 & 16 & 24 & 32 \end{array}\right). \tag{B.2}$$

The 2×2 partitions help to visualize the four modes of the resulting tensor: stepping from a partition to the one below increments the index of the first mode; stepping from a partition to the one on its right increments the index of the second mode; stepping down a row, within the same partition, increments the index of the third mode; stepping rightwards by a column, within the same partition, increments the index of the fourth mode. The symbol $^{\otimes R}$ denotes the Rth outer power of a tensor; that is, $\mathbf{A}^{\otimes R} = \overbrace{\mathbf{A} \otimes \mathbf{A} \otimes \cdots \otimes \mathbf{A}}^{R}$.

The *Khatri-Rao product* $\odot$ is the "matching columnwise" Kronecker product of matrices. The Khatri-Rao product of a matrix of size $I \times N$ and a matrix of size $J \times N$ is a matrix of size $IJ \times N$ (which may be interpreted as a tensor of size $I \times J \times N$). If $\mathbf{C} = \mathbf{A} \odot \mathbf{B}$, then $c[i, j, n] = a[i, n]\, b[j, n]$. This can be naturally extended to successive Khatri-Rao products of matrices: if $\mathbf{F} = \mathbf{A} \odot \mathbf{B} \odot \cdots \odot \mathbf{D}$, then $f[i, j, \ldots, \ell, n] = a[i, n]\, b[j, n] \cdots d[\ell, n]$ (the rows of the matrices, indexed here by n, must have the same dimension).[1] For example,

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \odot \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} = \left(\begin{array}{cc} 1{\cdot}5 & 1{\cdot}6 \\ 3{\cdot}7 & 3{\cdot}8 \\ \hline 2{\cdot}5 & 2{\cdot}6 \\ 4{\cdot}7 & 4{\cdot}8 \end{array}\right) = \left(\begin{array}{cc} 5 & 6 \\ 21 & 24 \\ \hline 10 & 12 \\ 28 & 32 \end{array}\right), \tag{B.3}$$

1 In Mathematica, this product can be written `Outer[Times, a, b, ..., d, 1]` where the final '`1`' specifies the level at which the outer product is calculated.

and

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \odot \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} \odot \begin{pmatrix} 9 & 11 \\ 10 & 12 \end{pmatrix}$$

$$= \left(\begin{array}{cc|cc} 1{\cdot}5{\cdot}9 & 3{\cdot}7{\cdot}11 & 1{\cdot}6{\cdot}9 & 3{\cdot}8{\cdot}11 \\ 1{\cdot}5{\cdot}10 & 3{\cdot}7{\cdot}12 & 1{\cdot}6{\cdot}10 & 3{\cdot}8{\cdot}12 \\ \hline 2{\cdot}5{\cdot}9 & 4{\cdot}7{\cdot}11 & 2{\cdot}6{\cdot}9 & 4{\cdot}8{\cdot}11 \\ 2{\cdot}5{\cdot}10 & 4{\cdot}7{\cdot}12 & 2{\cdot}6{\cdot}10 & 4{\cdot}8{\cdot}12 \end{array}\right)$$

$$= \left(\begin{array}{cc|cc} 45 & 231 & 54 & 264 \\ 50 & 252 & 60 & 288 \\ \hline 90 & 308 & 108 & 352 \\ 100 & 336 & 120 & 384 \end{array}\right). \tag{B.4}$$

As before, the partitions indicate the resulting tensors' modes. The symbol $\odot^R$ denotes the Rth Khatri-Rao power.

The *inner (dot) product* $\bullet$ is like the tensor product but additionally contracts (sums over the product of) the last index of the first tensor with the first index of the second tensor: if $\mathbf{C} = \mathbf{A} \bullet \mathbf{B}$, then $c[\ldots, i, j, \ell, m, \ldots] = \sum_k a[\ldots, i, j, k]\, b[k, \ell, m, \ldots]$ (the inner two modes of $\mathbf{A}$ and $\mathbf{B}$, indexed here by k, must have the same dimension). For an order-R tensor and an order-S tensor, this results in an order-$(R + S - 2)$ tensor. For example,

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \bullet \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} = 1{\cdot}4 + 2{\cdot}5 + 3{\cdot}6 = 32, \tag{B.5}$$

and

$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \bullet \begin{pmatrix} 5 & 7 \\ 6 & 8 \end{pmatrix} = \begin{pmatrix} 1{\cdot}5{+}3{\cdot}6 & 1{\cdot}7{+}3{\cdot}8 \\ 2{\cdot}5{+}4{\cdot}6 & 2{\cdot}7{+}4{\cdot}8 \end{pmatrix}$$

$$= \begin{pmatrix} 23 & 31 \\ 34 & 46 \end{pmatrix}. \tag{B.6}$$

C

COMPUTATIONAL SIMPLIFICATION OF EXPECTATION TENSORS

The general form of the expectation tensors is, as shown in Section 3.4.4,

$$x_{\mathrm{e}}[j_1, j_2, \ldots, j_R] = \sum_{\substack{(i_1,\ldots,i_R)\in\mathcal{I}^R:\\ i_n\neq i_p}} \prod_{r=1}^{R} x[i_r, j_r]\,, \tag{C.1}$$

which can be written in tensor notation as

$$\mathbf{X}_{JR} = \left(\left(\mathbf{1}_{JR} \otimes \mathbf{E}_{IR}\right) \circ \mathbf{X}^{\otimes R}_{\langle R+1,1,R+2,2,\ldots,\ldots,R+R,R\rangle}\right) \overset{R}{\bullet} \mathbf{1}_{IR}\,, \tag{C.2}$$

where $\mathbf{1}_{JR} \in \mathbb{R}^{J^R}$ is a tensor with R modes, each of dimension J, all of whose elements are ones, the $\overset{R}{\bullet}$ inner product with $\mathbf{1}_{I^R}$ represents R successive inner products with $\mathbf{1}_I$, and $\mathbf{E}_{dR}$ is constructed with elements

$$e[i_1, i_2, \ldots, i_R] = \begin{cases} 0 & \text{if } i_n = i_p \\ 1 & \text{otherwise.} \end{cases} \tag{C.3}$$

To understand the construction in (C.2), observe that the outer product $\mathbf{1}_{JR} \otimes \mathbf{E}_{IR}$ extends the tensor of nonrepeated indices into R additional modes, each of dimension J. Since $\mathbf{X}$ is an $I \times J$ matrix, $\mathbf{X}^{\otimes R} \in \mathbb{R}^{I\times J\times I\times J\times\cdots\times I\times J}$ is of order $2R$. The index permutation reshapes $\mathbf{X}^{\otimes R}$ into an element of $\mathbb{R}^{J^R\times I^R}$. The Hadamard product with the permuted $\mathbf{X}^{\otimes R}$, therefore, sets all entries occurring at locations with repeated indices to zero. These are precisely the entries that are excluded from the summation (C.1). The Rth inner product then sums

over the R different I-dimensional modes to collapse to the desired tensor in $\mathbb{R}^{J^R}$.

The expression takes this form due to the constraints on which index values are summed over. Both forms (C.1) and (C.2) are cumbersome to calculate directly. Were there no constraint on which indices in (C.1) are summed over, (C.2) would take the form

$$\mathbf{X}^{\otimes R}_{\langle R+1,1,R+2,2,\ldots,\ldots,R+R,R\rangle} \overset{R}{\bullet} \mathbf{1}_{IR} \,. \tag{C.4}$$

This requires $(IJ)^R$ multiplications, but can be reduced to J^R multiplications by rearranging it to

$$(\mathbf{1}_I'\mathbf{X})^{\otimes R} \,. \tag{C.5}$$

This suggests an alternative way of calculating (C.2), to sum all of the terms and then subtract the terms that should be excluded.

For example, consider the $R = 2$ case. The unconstrained term is $(\mathbf{1}_I'\mathbf{X})^{\otimes 2}$ and the term corresponding to the repeated indices is $(\mathbf{X}' \odot \mathbf{X}') \bullet 1_I$, which simplifies to $\mathbf{X}'\mathbf{X}$. Hence Equation (3.15) of the main text can be written

$$\mathbf{X}_{\mathrm{e}}^{(2)} = (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) - (\mathbf{X}'\mathbf{X}) \,. \tag{C.6}$$

The process for $R = 3$ is similar. The unconstrained term is $(\mathbf{1}_I'\mathbf{X})^{\otimes 3}$. There are three terms corresponding to the $i = j$ constraint, the $j = k$ constraint and the $i = k$ constraint, each is equal to one of the transpositions of $(\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})$. These have now subtracted out the $i = j = k$ constraint three times, and so $\mathbf{X}'^{\odot 3} \bullet 1_I$ must be added back in twice to

compensate. Accordingly, Equation (12) of the main text can be rewritten

$$\begin{aligned}\mathbf{X}_{\mathrm{e}}^{(3)} = (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,2,3\rangle} \\ - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 2,1,3\rangle} - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 3,1,2\rangle} \\ + 2\,(\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\,. \end{aligned} \tag{C.7}$$

An analogous procedure can be followed for any value of R, though this becomes increasingly difficult because the number of terms grows as $R!$. Each term represents a unique minimal set of different index constraints. For example, one term $\mathcal{A}$ might have the index constraints $i_1 = i_2$ and $i_3 = i_4$. Another term $\mathcal{B}$ might have no constraint on i_1 but have $i_2 = i_3 = i_4$. When the indices are ordered sequentially, the term can be calculated by writing each constraint as a subterm of the form

$$\mathbf{X}'^{\odot c} \bullet \mathbf{1}_I\,, \tag{C.8}$$

where c is the number of indices in that constraint, and then taking the outer product of the different subterms. For instance, with index constraints $\mathcal{A}$, (C.8) is

$$\Big((\mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big) \otimes \Big((\mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\,,$$

which simplifies to

$$(\mathbf{X}'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\ .$$

With index constraints $\mathcal{B}$, (C.8) is

$$(\mathbf{X}' \bullet \mathbf{1}_I) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\,,$$

which simplifies to $(\mathbf{1}_I'\mathbf{X}) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)$. The permutation of the indices in the constraints of a term is given by the corresponding permutation of that term's tensor. For example, the term with

constraints $i_1 = i_3$ and $i_2 = i_4$ (a permutation of $\mathcal{A}$) is represented by $\Big((\mathbf{X}'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,3,2,4\rangle}$ while the term with constraints i_2 and $i_1 = i_3 = i_4$ (a permutation of $\mathcal{B}$) is represented by $\Big((\mathbf{1}_I'\mathbf{X}) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\Big)_{\langle 2,1,3,4\rangle}$.

For example, the $R! = 24$ terms for the $R = 4$ case can be written

$$
\begin{aligned}
\mathbf{X}_{\mathrm{e}}^{(4)} = {} & (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,2,3,4\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,3,2,4\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,4,2,3\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 2,3,1,4\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 2,4,1,3\rangle} \\
& - \Big((\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{1}_I'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 3,4,1,2\rangle} \\
& + 2\Big((\mathbf{1}_I'\mathbf{X}) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\Big)_{\langle 1,2,3,4\rangle} \\
& + 2\Big((\mathbf{1}_I'\mathbf{X}) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\Big)_{\langle 2,1,3,4\rangle} \\
& + 2\Big((\mathbf{1}_I'\mathbf{X}) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\Big)_{\langle 3,1,2,4\rangle} \\
& + 2\Big((\mathbf{1}_I'\mathbf{X}) \otimes \Big((\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\Big)\Big)_{\langle 4,1,2,3\rangle} \\
& + \Big((\mathbf{X}'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,2,3,4\rangle} \\
& + \Big((\mathbf{X}'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,3,2,4\rangle} \\
& + \Big((\mathbf{X}'\mathbf{X}) \otimes (\mathbf{X}'\mathbf{X})\Big)_{\langle 1,4,2,3\rangle} \\
& - 6\,(\mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}' \odot \mathbf{X}') \bullet \mathbf{1}_I\,. \qquad \text{(C.9)}
\end{aligned}
$$

While expressions like (C.7) and (C.9) are harder to visualize than the more compact form (C.2), they can be calculated more efficiently: the unsimplified form has $O\Big(\big(IJ\big)^R\Big)$ multiplications, the simplified form

has $O\Big(I\big(J^R\big)\Big)$—a ratio of $1 : I^{R-1}$. Such simplifications are key in being able to calculate the practical examples of Section 3.6, of the main text, some of which use large values for I (102 in Ex. 3.6.2, and 19 in Ex. 3.6.3).

D

FORMAL SPECIFICATION OF THE MELODIC AFFINITY MODEL

In this appendix, I provide a full mathematical definition of the models and the form of the empirical data, that are presented in Chapter 4. Each of the mathematical steps are described in a more verbal manner in Sections 4.1.1, 4.1.2, and 4.3.2. The MATLAB function `Affinity_model_final.m` that implements these equations can be downloaded from `http://www.dynamictonality.com/melodic_affinity_files/`.

In the experiment, there were 44 participants, and each participant listened to 60 different stochastically generated melodies. The melodies were in one of eleven different *microtonal tunings* (the pitch intervals were different to those found in standard Western music). Furthermore, there were eleven different *timbres* used—each timbre had differently tuned partials (overtones or frequency components). The microtonal scales are indexed by $m \in \{1, 2, \ldots, 11\}$, and the timbres are indexed by $n \in \{1, 2, \ldots, 11\}$.

Each observation involved the participant listening to a single melody in tuning m played with two different timbres n_1 and n_2. The participant chose the timbre in which the melody's notes had the greater *affinity* (fitted the best). In all, 110 different stimuli (i.e., different values of the tuple (m, n_1, n_2)) were tested, but each participant listened to a randomly selected (uniform distribution without replacement) subset of 60 of these. A choice of n_1 was coded 1, a choice of n_2 was coded 0, and missing data was coded NaN: this results in a data matrix $\mathbf{Y} \in \{0, 1, \mathrm{NaN}\}^{110\times 44}$, each column containing 50 randomly located NaNs.

For timbre n, let $\mathbf{a}(n) \in \mathbb{Z}^{12}$ be a vector of the *pitches* (log-frequencies) of the first twelve partials indexed by i. The units of pitch are cents (1200th of an octave) above the first partial—the pitch of the first partial $a_1(n)$ is, therefore, always 0.

Let $\mathbf{w} \in \mathbb{R}^{12}$ be a vector of weights for the above 12 partials. The weights are the same, irrespective of the timbre n. The elements of $\mathbf{w}$ are also indexed by i, and their values are parameterized by a *roll-off* value $\rho \in \mathbb{R}$, so that

$$w_i(\rho) = i^{-\rho} \qquad i = 1, \ldots, 12\,. \tag{D.1}$$

This means that when $\rho = 0$, all partials have a weight of 1; as ρ increases, the weights of the higher partials are reduced.

The partials (their pitches and weights) are embedded in a cents domain *indicator vector* $\mathbf{b}(n;\rho) \in \mathbb{R}^{6000}$ whose elements are indexed by j:

$$b_j(n;\rho) = \sum_{i=1}^{12} w_i(\rho)\, \delta\Big[j - 1 - a_i(n)\Big] \qquad j = 1, \ldots, 6000\,, \tag{D.2}$$

where $\delta[z]$ is the Kronecker delta function, which equals 1 when $z = 0$, and equals 0 when $z \neq 0$. This equation means that the vector $\mathbf{b}(n;\rho)$ is all zeros except for twelve elements: for $i = 1$ to 12, its $(f_i(n) + 1)$th element has a value of w_i.

The twelve delta 'spikes' in $\mathbf{b}(n;\rho)$ are 'smeared' by non-circular convolution with a discrete Gaussian kernel $\mathbf{c}(\sigma)$ parameterized with a standard deviation $\sigma \in [0, \infty)$ to give a *spectral pitch vector* $\mathbf{d}(n;\rho,\sigma) \in \mathbb{R}^{6000}$, which is indexed by k:

$$d_k(n;\rho,\sigma) = \sum_{j=1}^{6000} b_j(n;\rho)\, c_{k-j+1}(\sigma) \qquad k = 1, \ldots, 6000\,, \tag{D.3}$$

where $c_z = 0$ if $z < 1$. For the sake of computational efficiency, the method used here differs slightly from that described in Ch. 3 where every partial is embedded into a separate vector and separately convolved before being summed into a final spectral pitch vector. The

spectral pitch vectors resulting from the two methods differ (slightly) only when the smeared pitches overlap in frequency. This would occur only when two partials are close in frequency, which is not the case for the timbres used in this experiment.

This vector is cross-correlated with the spectral pitch vector of a *harmonic template* $\mathbf{d}(\textsc{hct}; \rho, \sigma)$ (one with partials whose frequencies are integer multiples of the first partial) to produce a *virtual pitch vector* $\mathbf{g}(n; \rho, \sigma) \in \mathbb{R}^{11999}$, which is indexed by ℓ :

$$g_\ell(n; \rho, \sigma) = \sum_{k=1}^{6000} d_k(\textsc{hct}; \rho, \sigma)\, d_{k+\ell-6000}(n; \rho, \sigma) \qquad \ell = 1, \ldots, 11999\,, \tag{D.4}$$

where $d_x = 0$ if $6000 < x < 1$.

The *harmonicity* $h(n; \rho, \sigma) \in \mathbb{R}$ of the spectral pitch vector $\mathbf{d}(n; \rho, \sigma)$ is given by the maximum value of the above cross-correlation:

$$h(n; \rho, \sigma) = \max_\ell \Big(\mathbf{g}(n; \rho, \sigma) \Big)\,. \tag{D.5}$$

The *spectral pitch similarity* $s(n, u) \in (0, 1)$ of two tones with timbre n making an interval of size u cents is given by

$$s(n, u; \rho, \sigma) = \frac{\sum\limits_k d_k(n; \rho, \sigma)\, d_{k-u}(n; \rho, \sigma)}{\sum\limits_k \Big(d_k(n; \rho, \sigma) \Big)^2}\,. \tag{D.6}$$

This equation gives the cosine of the angle between the vector $\mathbf{d}(n; \rho, \sigma)$ and a transformation of itself that is shifted u elements to the right, and serves as a similarity measure between them.

The probability of an interval of size u cents occurring between two successive tones, given microtonal tuning m, is denoted $p_U(u \mid m)$. As described in Section 4.2.2, this probability distribution has eight independent parameters, which were constant across all stimuli. This

implies that the *expected spectral pitch similarity* $\overline{s}(m, n; \rho, \sigma)$ of successive tones with timbre n, given a microtonal tuning of m, is

$$\begin{aligned}\overline{s}(m, n; \rho, \sigma) &= E_u[s(n, u; \rho, \sigma) \mid m] \\ &= \sum_u p_U(u \mid m)\, s(n, u; \rho, \sigma)\,. \end{aligned} \tag{D.7}$$

Similarly, the *virtual pitch similarity* $v(n, u) \in (0, 1)$ of two tones with timbre n making an interval of size u cents is given by

$$v(n, u; \rho, \sigma) = \frac{\sum\limits_k g_k(n; \rho, \sigma)\, g_{k-u}(n; \rho, \sigma)}{\sum\limits_k \Big(g_k(n; \rho, \sigma)\Big)^2}\,, \tag{D.8}$$

and the *expected virtual pitch similarity* $\overline{v}(m, n; \rho, \sigma)$ of successive tones with timbre n, given a microtonal tuning of m, is

$$\begin{aligned}\overline{v}(m, n; \rho, \sigma) &= E_u[v(n, u; \rho, \sigma) \mid m] \\ &= \sum_u p_U(u \mid m)\, v(n, u; \rho, \sigma)\,. \end{aligned} \tag{D.9}$$

From (D.5), (D.7), and (D.9), I construct three predictors for the probability of choosing timbre n_1 given (m, n_1, n_2):

$$f_{\mathrm{S}}(m, n_1, n_2; \sigma, \rho) = \overline{s}(m, n_1; \rho, \sigma) - \overline{s}(m, n_2; \rho, \sigma) \tag{D.10}$$

$$f_{\mathrm{V}}(m, n_1, n_2; \sigma, \rho) = \overline{v}(m, n_1; \rho, \sigma) - \overline{v}(m, n_2; \rho, \sigma) \tag{D.11}$$

$$f_{\mathrm{H}}(n_1, n_2; \sigma, \rho) = h(n_1; \rho, \sigma) - h(n_2; \rho, \sigma)\,. \tag{D.12}$$

Three models, indexed by i, of the experimentally obtained data—the probabilities of choosing timbre n_1 for 110 different values of

(m, n_1, n_2)—were created from combinations of these predictors and differing parameterizations:

$$\begin{aligned}
\text{model}\, i &= P\Big(Y = 1 \mid m, n_1, n_2; i\Big) \\
&= \frac{1}{1 + \mathrm{e}^{-z_i}}\text{, where} \\
z_1 &= \beta_1 f_{\mathrm{S}}(m, n_1, n_2; \rho_{\mathrm{S}}, \sigma_{\mathrm{S}}) + \beta_2 f_{\mathrm{H}}(n_1, n_2; \rho_{\mathrm{H}}, \sigma_{\mathrm{H}}) \\
z_2 &= \beta_1 f_{\mathrm{S}}(m, n_1, n_2; \rho, \sigma) + \beta_2 f_{\mathrm{H}}(n_1, n_2; \rho, \sigma) \\
z_3 &= \beta_1 f_{\mathrm{V}}(m, n_1, n_2; \rho, \sigma) + \beta_2 f_{\mathrm{H}}(n_1, n_2; \rho, \sigma)\,. \qquad \text{(D.13)}
\end{aligned}$$

The overall form is a logistic model because the data are probabilities. A constant term is not used because (D.10–D.12) imply that if $n_1 = n_2$, then $f_{\mathrm{S}} = f_{\mathrm{V}} = f_{\mathrm{H}} = 0$.

E

CROSS-VALIDATION CORRELATION AND ROOT MEAN SQUARED ERROR

For each of the models discussed in Chapter 5, I performed twenty runs of 12-fold cross-validation of the models. Each of the twenty runs utilizes a different 12-fold partition of the probe tone data, each fold containing 2 samples. Within each run, one fold is removed and denoted the *validation set*; the remaining 11 folds are aggregated and denoted the *training set*. The parameters of the model are optimized to minimize the sum of squared errors between the model's predictions and the 22 samples in the training set. Cross-validation statistics, which measure the fit of the predictions to the validation set, are then calculated. This whole process is done for all twelve folds and this constitutes a single run of the 12-fold cross-validation. The same process is used for all twenty runs of the 12-fold cross-validation—each run using a different 12-fold partition of the data. The cross-validation statistics are averaged over all twelve folds in all twenty runs.

More formally: Let the data set of I samples be partitioned into K folds (the probe tone data comprise 24 values, so $I = 24$, and I use 12-fold cross-validation, so $K = 12$). Let $k[i]$ be the fold of the data containing the ith sample. The cross-validation is repeated, each time with a different K-fold partition, a total of J times. The cross-validation correlation of the jth run of the cross-validation is given by

$$r_{\mathrm{CV}}[j] = 1 - \sqrt{\sum_{i=1}^{I} (y_i - \widehat{y}_i^{\,\backslash k[i]})^2 \Big/ \sum_{i=1}^{I} (y_i - \overline{y})^2}\,, \tag{E.1}$$

where $\widehat{y}_i^{\,\backslash k[i]}$ denotes the fitted value for the ith sample returned by the model estimated with the $k[i]$th fold of the data removed, and $\overline{y}$ is the mean of all the sample values y_i. The final cross-validation correlation

statistic is the mean over the J runs of the cross-validation (in our analysis, $J = 20$):

$$r_{\mathrm{CV}} = \frac{1}{J} \sum_{j=1}^{J} r_{\mathrm{CV}}[j] \,. \tag{E.2}$$

The root mean squared error of the jth run of the cross-validation is given by

$$\mathrm{RMSECV}[j] = \sqrt{\frac{1}{I} \sum_{i=1}^{I} (y_i - \widehat{y}_i^{\setminus k[i]})^2} \,, \tag{E.3}$$

where $\widehat{y}_i^{\setminus k[i]}$ denotes the fitted value for the ith sample returned by the model estimated with the $k[i]$th fold of the data removed. The final root mean squared error of the cross-validation statistic is the mean over the J runs of the cross-validation:

$$\mathrm{RMSECV} = \frac{1}{J} \sum_{j=1}^{J} \mathrm{RMSECV}[j] \,. \tag{E.4}$$

F

FORMAL SPECIFICATION OF THE PROBE TONE DATA MODEL

In this appendix, I give a formal mathematical specification of the probe tone model described in Chapter 5. The techniques used are based on those introduced in Chapter 3. The MATLAB routines that embody these routines can be downloaded from `http://www.dynamictonality.com/probe_tone_files/`.

Let a chord comprising M tones, each of which contains N partials, be represented by the matrix $\mathbf{X}_{\mathrm{f}} \in \mathbb{R}^{M \times N}$. Each row of $\mathbf{X}_{\mathrm{f}}$ represents a tone in the chord, and each of element of the row is the frequency of a partial of that tone. In our model, I use the first twelve partials (so $N = 12$); this means that, if $\mathbf{X}_{\mathrm{f}}$ is a three-tone chord, it will be a 3×12 matrix.

The first step is to convert the partials' frequencies into pitch class cents values:

$$x_{\mathrm{pc}}[m,n] = 1200\lfloor \log_2(x_{\mathrm{f}}[m,n]/x_{\mathrm{ref}}) \rceil \bmod 1200\,, \tag{F.1}$$

where $\lfloor \cdot \rceil$ is the nearest integer function, and x_{ref} is an arbitrary reference frequency (e.g., the frequency of middle C). These values are then collected into a single *pitch class vector* denoted $\tilde{\mathbf{x}}_{\mathrm{pc}} \in \mathbb{Z}^{12M}$ indexed by i such that $x_{\mathrm{pc}}[m,n] \mapsto \tilde{x}_{\mathrm{pc}}[i]$, where $i = (m-1)N + n$.

Let each of the partials have an associated weight $x_{\mathrm{w}}[m,n]$, which represents their *salience*, or probability of being heard. I test three models (a, b, and c). Given model ℓ, where $\ell \in \{a,b,c\}$ denotes the model, the saliences of the tonic triad are parameterized by a *roll-off* value $\rho \in \mathbb{R}$, and a *chord-degree weighting* value $\omega \in [0,1]$, so that

$$\omega^{[m \notin \mathcal{R}_\ell]} x_{\mathrm{w}}[m,n] = n^{-\rho} \qquad m = 1,\ldots,M, \text{ and } n = 1,\ldots,12\,, \tag{F.2}$$

where $[m \notin \mathcal{R}_\ell]$ denotes an indicator function that equals 0 when tone m is member of the set $\mathcal{R}_\ell$ of tones classed as chord roots in model ℓ, and is otherwise 1. In model a, all tones are classed as roots, hence all tones have a chord-degree weighting of 1; in model b, only the conventional roots of the major and minor triads are classed as roots (i.e., pitch class C in the chord Cmaj or Cmin), all other tones have a chord degree weighting of ω; in model c, the third of the minor triad is also classed as a root (e.g., E♭ in Cmin), the remaining tones have a chord degree weighting of ω. Ignoring the chord degree weighting value, Equation (F.2) means that when $\rho = 0$, all partials of a tone m have a weight of 1; as ρ increases, the weights of its higher partials are reduced. These values are collected into a single *weighting vector* $\widetilde{\mathbf{x}}_{\mathrm{w}} \in \mathbb{R}^{12M}$ also indexed by i such that $x_{\mathrm{w}}[m, n] \mapsto \widetilde{x}_{\mathrm{w}}[i]$, where $i = (m-1)N + n$ (the precise method used to reshape the matrix into vector form is unimportant so long as it matches that used for the pitch class vector).

The partials (their pitch classes and weights in $\widetilde{\mathbf{x}}_{\mathrm{pc}}$ and $\widetilde{\mathbf{x}}_{\mathrm{w}}$) are embedded in a *spectral pitch class salience matrix* $\mathbf{X}_{\mathrm{pcs}} \in \mathbb{R}^{12N \times 1200}$ indexed by i and j:

$$x_{\mathrm{pcs}}[i, j] = \widetilde{x}_{\mathrm{w}}[i]\, \delta\Big[j - \widetilde{x}_{\mathrm{pc}}[i]\Big] \qquad i = 1, \ldots, 12N, \text{ and } j = 0, \ldots, 1199\,, \tag{F.3}$$

where $\delta[z]$ is the Kronecker delta function, which equals 1 when $z = 0$, and equals 0 when $z \neq 0$. This equation means that the matrix $\mathbf{X}_{\mathrm{pcs}}$ is all zeros except for $12N$ elements, and each element indicates the salience $x_{\mathrm{pcs}}[i, j]$ of partial i at pitch j.

To model the uncertainty of pitch perception, these $12N$ delta "spikes" are "smeared" by circular convolution with a discrete Gaussian kernel $\mathbf{g}$, which is also indexed by j, and is parameterized with a *smoothing* standard deviation $\sigma \in [0, \infty)$ to give a *spectral pitch class response matrix* $\mathbf{X}_{\mathrm{pcr}} \in \mathbb{R}^{12N \times 1200}$, which is indexed by i and k:

$$\mathbf{x}_{\mathrm{pcr}}[i] = \mathbf{x}_{\mathrm{pcs}}[i] * \mathbf{g}\,, \tag{F.4}$$

where $\mathbf{x}_{\mathrm{pcr}}[i]$ is the ith row of $\mathbf{X}_{\mathrm{pcr}}$, and $*$ denotes circular convolution over the period of 1200 cents; that is,

$$x_{\mathrm{pcr}}[i,k] = \sum_{j=0}^{1199} x_{\mathrm{pcs}}[i,j]\, g[(k-j) \bmod 1200]$$
$$i = 1,\ldots,12N, \text{ and } k = 0,\ldots,1199\,. \tag{F.5}$$

In this implementation, I make use of the circular convolution theorem, which allows (F.4) to be calculated efficiently with fast Fourier transforms; that is, $\mathbf{f} * \mathbf{g} = \mathcal{F}^{-1}\Big(\mathcal{F}(\mathbf{f}) \circ \mathcal{F}(\mathbf{g})\Big)$, where $*$ is circular convolution, $\mathcal{F}$ denotes the Fourier transform, $\circ$ is the Hadamard (elementwise) product, and $\mathbf{f}$ stands for $\mathbf{x}_{\mathrm{pcs}}[i]$.

Equation (F.4) can be interpreted as adding random noise (with a Gaussian distribution) to the original pitch classes in $\mathbf{X}_{\mathrm{pcs}}$, thereby simulating perceptual pitch uncertainty. The standard deviation of the Gaussian distribution σ models the pitch difference limen (just noticeable difference) (App. A). In laboratory experiments with sine waves, the pitch difference limen is approximately 3 cents in the central range of frequency (Moore, 1973; Moore et al., 1984). We would expect the pitch difference limen in the more distracting setting of listening to music to be somewhat wider. Indeed, the value of σ was optimized—with respect to the probe tone data—at approximately 6 cents.

Each element $x_{\mathrm{pcr}}[i,k]$ of this matrix models the probability of the ith partial in $\mathbf{x}_{\mathrm{p}}$ being heard at pitch class k. In order to summarize the responses to all the pitches I take the column sum, which gives a vector of the expected numbers of partials heard at pitch class k. This 1200-element row vector is denoted a *spectral pitch class vector* $\mathbf{x}$:

$$\mathbf{x} = \mathbf{1}'\mathbf{X}_{\mathrm{pcr}}\,, \tag{F.6}$$

where $\mathbf{1}$ is a column vector of $12N$ ones, and $'$ denotes matrix transposition, so $\mathbf{1}'$ is a row vector of $12N$ ones.

The spectral pitch similarity of two such vectors $\mathbf{x}$ and $\mathbf{y}$ is given by any standard similarity measure. I choose the cosine:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}\mathbf{y}'}{\sqrt{(\mathbf{x}\mathbf{x}')(\mathbf{y}\mathbf{y}')}} \,. \tag{F.7}$$

This similarity value falls between 0 and 1, where 1 implies the two vectors are parallel, and 0 implies they are orthogonal.

I use this model to establish the similarities of a variety of probes with respect to a context. Let the context be represented by the spectral pitch class vector $\mathbf{x}$, and let the P different probes $\mathbf{y}_p$ be collected into a matrix of spectral pitch class vectors denoted $\mathbf{Y} \in \mathbb{R}^{P \times 1200}$. The column vector of P similarities between each of the probes and the context is then denoted $\mathbf{s}(\mathbf{x}, \mathbf{Y}) \in \mathbb{R}^P$. For example, the context may be a major triad built from HCTs and the probes may be single HCTs at the twelve chromatic pitches. In this case, the thirty-six harmonics from the context (12 partials for each of the three different chord tones) are embedded into a single spectral pitch class vector $\mathbf{x}$, as described in (F.1–F.6). Each of the twelve differently pitched probe tones' 12 harmonics are embedded into twelve spectral pitch class vectors $\mathbf{y}_p$. The similarities of the context and the twelve probes are calculated—as described in (F.7)—to give the vector of their similarities $\mathbf{s}(\mathbf{x}, \mathbf{Y})$.

Models a, b, and c can now be summarized in mathematical form: Let the vector of probe tone data for both contexts be denoted $\mathbf{d} \in \mathbb{R}^{24}$; let the vector of associated modelled similarities be denoted $\mathbf{s}(\mathbf{x}, \mathbf{Y}; \rho, \sigma, \omega, \ell) \in \mathbb{R}^{24}$, where ρ, σ, ω are the roll-off, smoothing, and chord degree weighting parameters discussed above, and $\ell \in \{a, b, c\}$ denotes the model; let $\mathbf{1}$ be a column vector of 24 ones;

$$\mathbf{d} = \alpha\mathbf{1} + \beta\mathbf{s}(\mathbf{x}, \mathbf{Y}; \rho, \sigma, \omega, \ell) + \boldsymbol{\epsilon} \,, \tag{F.8}$$

where α and β are the linear intercept and slope parameters, and $\boldsymbol{\epsilon}$ is a vector of 24 unobserved errors that captures unmodelled effects or random noise.

Each model's parameter values were optimized, iteratively, to minimize the sum of squared residuals between the model's predictions and the empirical data; that is, the optimized parameter values for model ℓ are given by

$$\begin{aligned}&(\widehat{\alpha}, \widehat{\beta}, \widehat{\rho}, \widehat{\sigma}, \widehat{\omega})[\ell] \\ &= \underset{\alpha,\beta,\rho,\sigma,\omega}{\arg\min} \left(\mathbf{d} - \alpha\mathbf{1} - \beta\mathbf{s}(\rho, \sigma, \omega, \ell)\right)' \left(\mathbf{d} - \alpha\mathbf{1} - \beta\mathbf{s}(\rho, \sigma, \omega, \ell)\right),\end{aligned} \tag{F.9}$$

where $\arg\min\ f(\boldsymbol{\theta})$ returns the value of $\boldsymbol{\theta}$ that minimizes the value of $f(\boldsymbol{\theta})$.